Leitlinien zum Management von Forschungsdaten: Sozialwissenschaftliche Umfragedaten

Jensen, Uwe

Monographie / monograph

gesis Leibniz-Institut für Sozialwissenschaften

TECHNICAL *Reports* 2012|07

Leitlinien zum Management von Forschungsdaten

Sozialwissenschaftliche Umfragedaten

Uwe Jensen

GESIS-Technical Reports 2012|07

Leitlinien zum Management von Forschungsdaten

Sozialwissenschaftliche Umfragedaten

Uwe Jensen

unter Mitarbeit von Evelyn Brislinger, Kristina Hauser, Brigitte Hausstein, Alexia Katsanidou, Dafina Kurti, Laurence Horton, Reiner Mauer, Meinhard Moschner, Markus Quandt, Astrid Recker, Natascha Schumann, Oliver Watteler, Wolfgang Zenk-Möltgen

GESIS – Leibniz-Institut für Sozialwissenschaften 2012

GESIS-Technical Reports

GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln
50667 Köln
Telefon: (0221) 476 94 - 0
Telefax: (0221) 476 94 - 199
E-Mail: uwe.jensen@gesis.org

ISSN: 1868-9043 (Print)
ISSN: 1868-9051 (Online)
Herausgeber,
Druck und Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Unter Sachsenhausen 6-8, 50667 Köln

Inhaltsverzeichnis

Verzeichnis der Abbildungen und Übersichten

Einleitung

Der verantwortungsvolle Umgang mit Forschungsdaten wird national und international mit der Erwartung verbunden, die Ergebnisse öffentlich geförderter Forschungsprojekte für weitere Analysen und Replikationen dauerhaft und global verfügbar zu machen (OECD 2007; Allianz 2010; EU 2010; NSF 2011). Das Thema Forschungsdatenmanagement hat auch für sozialwissenschaftliche Forschung in Deutschland eine besondere Bedeutung gewonnen, seitdem die DFG ihre Förderung mit der Beschreibung von Maßnahmen zum Datenmanagement in datenerzeugenden Forschungsprojekten verknüpft (DFG 2011; DFG 2012b; DFG 2012c).

Was dabei unter Daten im Zuge eines zeitgemäßen Managements von Forschungsdaten zu verstehen ist, wird je nach Forschungskontext und Wissenschaftsdisziplin unterschiedlich begrifflich beschrieben. In den „Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten" der DFG wird in Empfehlung Nr. 1 von Daten gesprochen,

- „die im Verlauf von Quellenforschungen, Experimenten, Messungen, Erhebungen oder Umfragen entstanden sind. Sie stellen die Grundlagen für die wissenschaftlichen Publikationen dar."
- Sie sind je nach Fachdisziplin „unterschiedlich zu definieren". Ob „bereits Rohdaten hierzu zählen oder ab welchem Grad der Aggregierung die Daten langfristig aufzubewahren sind". ist von den beteiligten Wissenschaftlern zu entscheiden (DFG 2009: 2).

Die Anforderungen an ein systematisches Management von Daten in einem Forschungsvorhaben sind vielfältig. Rein technische Maßnahmen der Datensicherung greifen zu kurz, wenn die Daten am Projektende für die weitere Nutzung gesichert und bereitgestellt werden sollen. Ein integriertes und aktives Management von Forschungsdaten verfolgt deshalb das Ziel, alle datenrelevanten Fragen vom Studien- und Methodendesign über die Datenaufbereitung bis hin zur Datenarchivierung in die Planung eines Forschungsprojektes einzubeziehen und zu dokumentieren. Ein strukturiertes Konzept zum Datenmanagement, das in die methodisch-theoretische Planung und Durchführung einer empirischen Untersuchung eingebettet ist, kann den Projektalltag erheblich entlasten. Es hilft zugleich, neuen Anforderungen von Förderorganisationen zum Umgang mit Forschungsdaten im Rahmen von Projektanträgen frühzeitig und effizient gerecht zu werden.

Ziel dieser Publikation ist es, Grundlagen des Managements sozialwissenschaftlicher Forschungsdaten, insbesondere von Umfragedaten, darzustellen. Dazu werden u. a. Standards, Leitlinien und Empfehlungen des Datenarchivs für Sozialwissenschaften zum Datenmanagement aufgegriffen, die in Kooperationen mit Forschungsprojekten und zur Archivierung und Erschließung quantitativer Daten der empirischen Sozialforschung genutzt werden. Die Publikation richtet sich an Forscherinnen und Forscher aus den Sozialwissenschaften und Interessierte aus anderen Disziplinen, die sich ausbildungs- oder forschungsbegleitend mit dem Thema Forschungsdatenmanagement befassen wollen, an Fragen standardisierter Datenaufbereitung und Datendokumentation interessiert sind oder Projektdaten pragmatisch für eine längerfristige Sicherung und Nutzung vorbereiten wollen.

Grundlegende Aspekte und Leitlinien, die beim Management von Forschungsdaten in Projektkontexten eine zentrale Rolle spielen, werden überblicksartig anhand von sieben Themen in Kapitel 1 behandelt.

Anforderungen an das Datenmanagement lassen sich dabei entlang der Projektphasen eines Forschungsdatenzyklus verorten (Abbildung 1, S. 12). Kernelemente, die auch bei der Erstellung von Datenmanagementplänen (vgl. Jensen 2011) beachtet werden sollten, werden anschließend in einer Checkliste zum Forschungsdatenmanagement (S. 19) zusammengefasst.

Die Liste erfasst anhand von unterschiedlichen Themenblöcken wichtige Fragestellungen zum Datenmanagement, um die spezifische Planung sozialwissenschaftlicher Datenprojekte zu unterstützen. Interessierte, die Informationen zu besonderen Aspekten suchen, finden hierzu Verweise auf weiterführende Kapitel und Abschnitte.

Einige der zentralen Themen des Datenmanagements werden in den folgenden Kapiteln vertieft.

Kapitel 2 behandelt Regeln, Abläufe und Konventionen in der Datenaufbereitung. Inhaltliche Aspekte des Datenmanagements werden im Zusammenhang der Datendefinition und Datenbereinigung quantitativer Daten der empirischen Sozialforschung beschrieben.

Im Mittelpunkt von Kapitel 3 stehen datenschutzrechtliche sowie logische und technische Aspekte der Organisation und Sicherung von Dateien und Dokumenten in einem Forschungsprojekt. Dabei wird auch auf ein Konzept zur Versionierung von Datensätzen eingegangen.

In Kapitel 4 sind Metadaten und Standards in sozialwissenschaftlichen Forschungskontexten das zentrale Thema. Zunächst wird auf relevante Standards (z. B. zur Demographie) sowie internationale Klassifikationen und Normen zur Erhebung bzw. Dokumentation von Daten hingewiesen. Im Zusammenhang mit Anforderungen an eine standardisierte Dokumentation von Studien und Daten wird anschließend der technische DDI-Standard vorgestellt. Mit dem Angebot des Kooperationsprojektes da|ra werden die Grundlagen und Möglichkeiten zur dauerhaften Zitation von Sozial- und Wirtschaftsdaten skizziert. Auch hier können Metadaten und kontrollierte Vokabulare auf Basis eines spezifischen Metadatenschemas standardisiert beschrieben und langfristig für die inhaltsreiche Erschließung zitierter Forschungsdaten eingesetzt werden.

Kapitel 5 widmet sich Fragen der Sicherung und der langfristigen Nutzung von Forschungsdaten, die am Ende des Forschungsprojektes vorliegen. Zunächst werden mögliche organisatorische Optionen und damit zusammenhängende Fragen der zeitlichen Dauer der Datensicherung und die Zugänglichkeit der Daten thematisiert. Anforderungen an Metadaten und Dokumentationen auf Studien- und Datenebene werden im Zusammenhang mit der Projektplanung zur nachhaltigen Archivierung und Bereitstellung der Forschungsdaten behandelt. Damit verbundene rechtliche Aspekte und die mögliche Regelung durch einen Archivierungsvertrag mit dem GESIS Datenarchiv werden in einem eigenen Abschnitt vorgestellt. Fragen zur Auswahl und Übergabe der zu sichernden Daten und Dokumente werden zum Schluss beschrieben.

In Kapitel 6 werden die einzelnen Angebote des GESIS Datenarchiv für Sozialwissenschaften für Datengeber vorgestellt.

Die im Text genannten Quellen, Referenzen und Beispiele, die aus unterschiedlicher Perspektive Materialien zum Management von Forschungsdaten bereitstellen, sind im Anhang zusammengestellt.

- Der Anhang A.1 führt nationale und internationale Datenquellen und Datenanbieter auf, die umfassend dokumentierte Daten für Sekundäranalysen bereitstellen.
- Der Anhang A.2 enthält die Literaturangaben sowie Verweise auf Förderrichtlinien, Materialien und Dokumentationen zum Datenmanagement, wie sie insbesondere von komplexen Forschungs- und Umfrageprogrammen erstellt werden.
- Im Anhang A.3 werden die Abschnitt 4.1 erwähnten sozialwissenschaftlich relevante Standards, Klassifikationen und Normen ausgewiesen.
- Anhang A.4 führt die Quellen zu technischen Metadatenstandards und Initiativen sowie Publikationen, Beispiele und Quellen aus Abschnitt 4.2 und 4.3 auf.

1 Themen des Forschungsdatenmanagements in der Projektplanung

Vor dem Hintergrund der jeweiligen Forschungsfragen und des Standes der Forschung werfen, der zu entwickelnde Forschungs- und Arbeitsplan, die Projektorganisation und nicht zuletzt die Kosten des Vorhabens eine Reihe praktischer Fragen bei der Planung eines Forschungsprojektes auf. In Ergänzung zu forschungslogischen und methodologischen Überlegungen sind auch eine Reihe von Fragen zum Management der Daten im gesamten Projektverlauf (vgl. Abbildung 1) zu berücksichtigen, wenn z. B. eine sozialwissenschaftliche Umfrage geplant wird.

Um welche Forschungsdaten es im Projekt geht, wie sie erhoben und für Analysen aufbereitet werden sollen oder welche Erfordernisse bei der Dokumentation der Daten zu berücksichtigen sind, all diese Aspekte sind bei der Planung des Forschungsvorhabens einzubeziehen und in einem Datenmanagementplan (DMP) zu dokumentieren (vgl. Jensen 2011). Der im DMP festgelegte Umgang mit den Forschungsdaten im Projektverlauf hängt dabei auch davon ab, wie die Daten sowohl während als auch am Ende des Projektes gesichert und weitergenutzt werden sollen. Fragen nach der Art und Dauer der Datenaufbewahrung sowie die Bereitstellung der Daten für die weitere Forschung rücken dann in den Vordergrund. Forscher sollen diese Aspekte des Datenmanagements vor Projektbeginn auch als einen wichtigen Teil in einem Antrag auf Forschungsförderung behandeln.

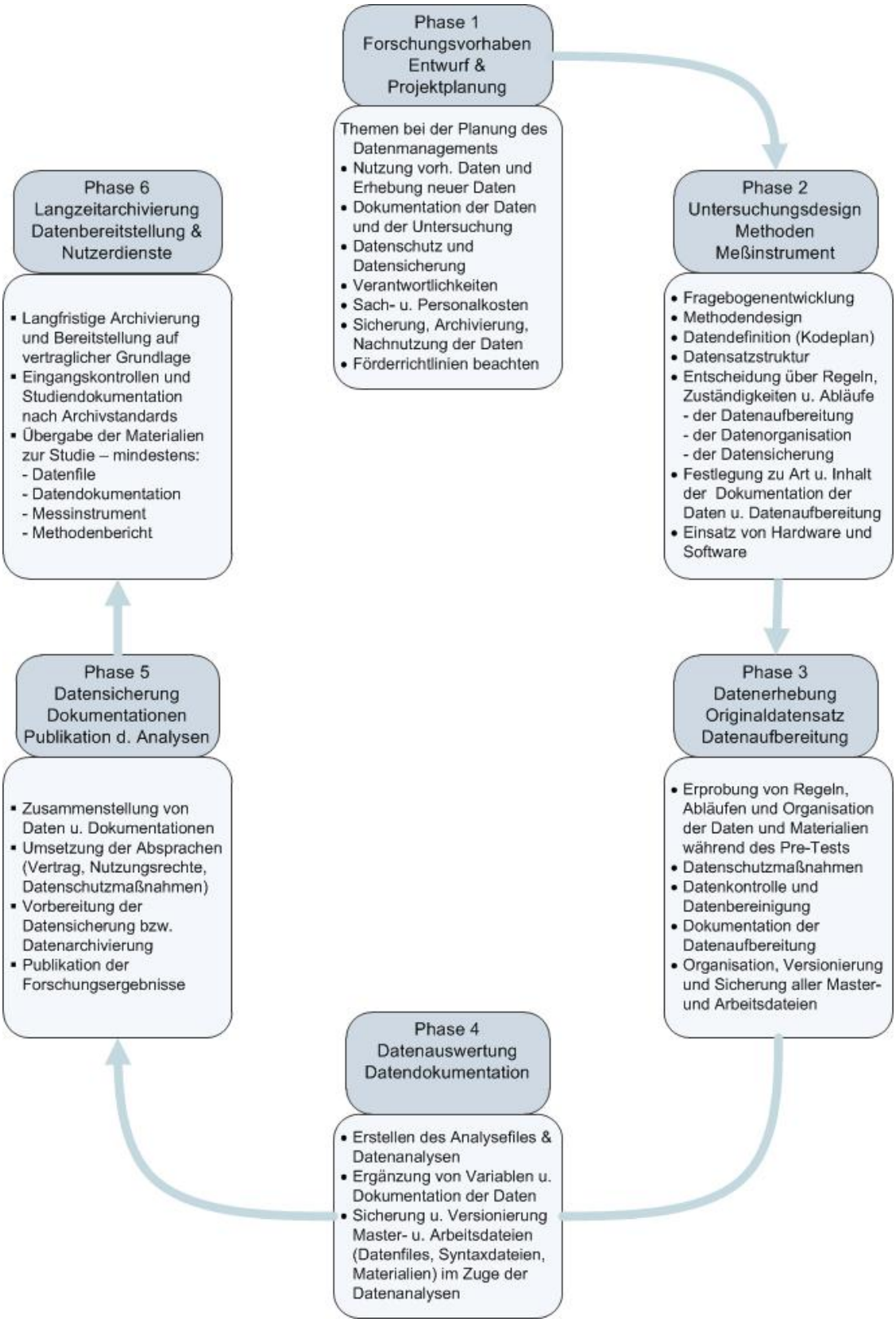
1.1 Exploration vorhandener Daten und Erhebung neuer Forschungsdaten

Datengenerierende Projekte sollten neben den üblichen Literaturrecherchen zum Stand der Forschung auch ermitteln, ob bzw. welche (themen- und domänenspezifischen) Datensätze bereits vorhanden sind. In diesem Kontext ist zu untersuchen, ob zur Beantwortung einer Fragestellung eine Datenerhebung überhaupt erforderlich ist bzw. ob vorhandene Daten in das geplante Datendesign des Forschungsprojektes einbezogen werden können.

Sekundäranalysen von bestehenden Datenbeständen stellen eine anerkannte Forschungsstrategie dar. Sie können der Vorbereitung eigener Erhebungen dienen (z. B. Prüfen von Fragenformulierungen) oder auch wissenschaftliche Analysen ermöglichen, ohne selbst Daten erheben zu müssen (z. B. als Vergleich von Datensätzen, die sich thematisch und methodisch gleichen). Es kann aber auch das methodische Forschungsinteresse eines Projektes sein, vorhandene empirische Daten mit modifizierter Fragestellung neu zusammenzustellen oder mit verbesserten Methoden zu reanalysieren und neu zu präsentieren. Neben klassischen Replikationsstudien oder Studien zur Kontrolle von Primärdaten können Projekte ihre Datenbasis mit Forschungsdaten aus unterschiedlichen Datenquellen erweitern oder verknüpfen (vgl. Häder 2010:129f; Kromrey 2009: 506f; Schnell, Hill, Esser 2011: 243f). Die Festlegung der Datenbasis durch das Forschungsprojekt stellt einen zentralen Aspekt dar, der bei der Formulierung eines Datenmanagementplans zu berücksichtigen ist. Auf Grundlage dieser Informationen kann dann die Notwendigkeit der projektspezifischen Datenerhebung oder die Auswertung vorhandener Daten bei Drittmittelanträgen begründet und in den aktuellen Erkenntnisstand eingeordnet werden.

Voraussetzung für Sekundäranalysen von Umfragedaten oder anderen Datentypen ist, dass entsprechende Daten auffindbar, verständlich dokumentiert und zugänglich sind. Datenkataloge und Serviceangebote der international vernetzten sozialwissenschaftlichen Datenarchive können genutzt werden, um Recherchen in Datenbeständen oder umfassende Dokumentationen von Umfragedaten durchzuführen und entsprechende Daten für das eigene Forschungsprojekt anzufordern. Je nach Forschungsdesign können auch die Daten der amtlichen Statistik sowie prozessproduzierte Daten aus Behörden und internationalen Organisation als Quellen für die angestrebten Datenanalysen genutzt werden (vgl. Anhang A.1).

Abbildung 1: Forschungsdatenmanagement - Anforderungen im Projektverlauf



1.2 Datenmanagement und Dokumentationen auf Studien- und Variablenebene

Wird eine eigene Datenerhebung im Projekt geplant, rücken die Fragen an die Datendokumentation und das Management der zu bearbeitenden Daten in den Vordergrund. Sie sind wiederum eng mit der methodischen Entwicklung des Erhebungsdesigns und der Durchführung der sozialwissenschaftlichen Untersuchung verknüpft. Dabei erfordern die Prozesse und Regeln zur Datendefinition, Datenaufbereitung und der anschließenden Datenanalysen sowie die Darstellung des Projektes und seiner Ergebnisse eine besondere Beachtung.

Um das Management der Forschungsdaten im Projekt zu planen, ist es hilfreich, die Definition und Struktur der Variablen und die Software für die Datenbearbeitung und Datenauswertung schon vor der Erhebung festzulegen.

Liegt der Fragebogen (-entwurf) vor und ist über die Erhebungstechnik entschieden, können auch die wesentlichen Standards und Codierregeln zur Datenaufbereitung sowie Leitlinien zur Datendokumentation verbindlich für den gesamten Datenworkflow festgelegt werden. Das frühzeitige Anlegen eines Codeplans, der Fragen und Variablen aufeinander bezieht, schafft eine Grundlage, um die erzeugten Daten sowie ihre Kontrolle, Bereinigung und Transformation im Projektverlauf angemessen und transparent zu dokumentieren (vgl. Kapitel 2).

Der Pretest des Fragebogens kann dann mit einem Probelauf zur Datenbearbeitung und Datendokumentation verknüpft werden. Dadurch lassen sich Schwachstellen und Schwierigkeiten frühzeitig erkennen. Relevante Prozesse und Leitlinien zum Umgang mit den Daten können dann vor der eigentlichen Datenerhebung und der anschließenden Datenaufbereitung angepasst und verbessert werden.

Weiterhin können Konventionen zur Datenorganisation (Formate, Namen, Versionierung von Dateien) und Datensicherung (Plattformen, Regeln der Datensicherung) frühzeitig bei der Planung des Datenmanagements in Betracht gezogen werden (vgl. Kapitel 3).

Die den Forschungsprozess begleitenden Maßnahmen zum Datenmanagement schaffen die Voraussetzungen für die Integration der Abläufe und Regeln zum Datenumgang und entsprechende Verantwortlichkeiten in die gesamte Projektorganisation. Situativ bedingte Ad hoc Planungen mit ihren Reibungen und Zeitverlusten lassen sich dadurch reduzieren.

In diesem Zusammenhang ist auch zu überlegen, welche Arten und Formen der Dokumentation notwendig und geeignet sind, um die relevanten Informationen auf Studien-, Methoden- und Datenebene systematisch im Projektverlauf zu erfassen, und - soweit geplant - mit der nachhaltigen Nutzung der Daten bereitzustellen (vgl. Abschnitt 5.2).

Inwieweit Standards und Regelwerke zur inhaltlichen bzw. technischen Dokumentation der Studie- und der Daten genutzt werden können, ist anhand verschiedener Kriterien zusätzlich im Projekt bzw. im Zuge von Beratungen durch erfahrene Kooperationspartner oder datenhaltende Einrichtungen zu prüfen. Darüber hinaus können sozialwissenschaftlich relevante Standards, Normen oder Klassifikationen berücksichtigt werden. Die standardisiert Erhebung und Dokumentation spezieller Daten von Befragten (z. B. sozio-demographische Merkmale, Berufe) erhöht zugleich das Potential für vergleichende Datenanalysen (vgl. Kapitel 4).

1.3 Regelungen zum Schutz personenbezogener Daten

Der Schutz der Untersuchungsperson sowie der Schutz vor einer missbräuchlichen Verwendung ihrer personenbezogenen Informationen stellen Kernaspekte des verantwortlichen Forschens im sozialwissenschaftlichen Umfeld dar. Ein Projekt, das personenbezogene Daten z. B. durch persönliche, schriftli-

che, telefonische oder andere Verfahren zu Forschungszwecken erhebt und verarbeitet, ist deshalb an die Regeln des Datenschutzes nach dem Bundesdatenschutzgesetz bzw. Länderregelungen zum Datenschutz gebunden. Dies gilt soweit kein bereichsspezifisches Gesetz besondere Regelungen zur Datenerhebung und –verarbeitung vorschreibt und damit dem Bundesdatenschutz vorgeht (Subsidiärprinzip). Dies ist z. B. das Zehnte Sozialgesetzbuch (SGB X), das etwa im Fall von medizinischen Daten greift oder das für die Daten des Mikrozensus zuständige Bundesstatistikgesetz (BStatG).

Personenbezogene Daten werden als „Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person" verstanden (BDSG § 3 Absatz 1). D. h., neben den konkreten Angaben einer befragten Person unterliegen insbesondere auch die Daten dem Datenschutz, die es indirekt ermöglichen könnten, die Identität einer Person festzustellen. Diese sog. individualisierbaren bzw. personenbeziehbaren Daten können etwa den Einzelangaben über persönliche oder sachliche Verhältnisse z. B. zu Beruf, Alter, Geschlecht und Wohnort entstammen.

Weiterhin gibt das Recht auf informationelle Selbstbestimmung dem Einzelnen das Recht, „über die Preisgabe und Verwendung seiner persönlichen Daten" selbst zu entscheiden (BVerfGE 65, 1). Dies setzt die Möglichkeit voraus, durch Einwilligung oder Verweigerung zu entscheiden, ob bzw. unter welchen Bedingungen die eignen personenbezogenen Daten erhoben und verarbeitet werden.

Dazu werden in den datenschutzrechtlichen Regelungen (z. B. § 4 BDSG und Länderregelungen) formale und inhaltliche Anforderungen an eine Einverständniserklärung festgelegt (vgl. Metschke, Wellbrock 2002: 25ff), die als Konzept des „informed consent" bekanntgeworden sind. Es regelt damit zentrale Grundrechte, die in der Beziehung zwischen Forscher („Forschungsfreiheit") und Befragten („informationelle Selbstbestimmung") zum Tragen kommen. Die informierte Einwilligung zur Datenerhebung und -verarbeitung personenbezogener Daten durch ein Forschungsvorhaben beruht auf den folgenden Grundsätzen des BDSG:

1. Informierte Einwilligung (informed consent) durch Aufklärung mit Hilfe von wesentlichen Informationen über das Forschungsvorhaben und den Zweck der geplanten Verarbeitung erhobener Daten. Weiterhin sind Betroffene u. a. über das Recht zu informieren, die Einwilligung in die Datenverarbeitung zu widerrufen.
2. Die informierte Einwilligung der Betroffenen erfolgt zweckgebunden durch Bezug auf ein konkret benanntes Forschungsvorhaben.
3. Die Einwilligung erfolgt freiwillig.
4. Die Einwilligung erfolgt in der Regel schriftlich. Die Schriftlichkeit wird generell in den Regelungen zum Datenschutz verlangt, da sie u. a. Schutz-, Garantie- und Beweiszwecken dient.

   Eine strikt schriftliche Einwilligungsform erzeugt aber Probleme für die sozialwissenschaftliche Forschung, z. B. bei Telefoninterviews oder durch erhöhte Verweigerungsraten bei face-to-face Befragungen. Um diesen besonderen Umständen Rechnung zu tragen, wurden auch Möglichkeiten der sozialwissenschaftlichen Datenerhebung auf Grundlage einer mündlichen Einwilligung des Betroffenen datenschutzrechtlich geregelt.

Muster datenschutzkonformer Einverständniserklärungen werden vom ADM (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V.) bereitgestellt (ADM, o. J. Vergleiche auch die Beispiele für Einwilligungserklärungen in Metschke, Wellbrock 2002: 55ff). Die Umstände, die unterschiedliche Einwilligungsformen im Rahmen einmaliger sozialwissenschaftlicher Erhebungen zulassen, beschreiben Metschke, Wellbrock (2002: 29f) sowie Häder (2010: 136f).

Einen vertiefenden Einblick in das Thema datenschutzrechtlicher und ethischer Fragen, mit denen sozialwissenschaftlich arbeitende Forscherinnen und Forscher bei der Projektplanung konfrontiert sein können, vermittelt die Publikation „Der Datenschutz in den Sozialwissenschaften" (Häder 2009). Im

Mittelpunkt des Beitrages stehen praktische Anforderungen des Datenschutzes, die während einzelner Phasen des Forschungsprozesses in der qualitativen empirischen Sozialforschung berücksichtigt werden sollten. Dies betrifft u. a. die Durchführung von Pretests, die Stichprobenziehung und die Feldarbeit sowie verschiedene Erhebungsverfahren zur Gewinnung personenbezogener Daten.

Im Vergleich zu einmalig durchgeführten Querschnittserhebungen sind bei Wiederholungs- und Folgebefragungen (Panelstudien) weit aufwendigere Maßnahmen zum Datenschutz bei der Verarbeitung personenbezogener Daten zu ergreifen. Dies betrifft sowohl die Daten aus der wachsenden Anzahl von Erhebungswellen als auch die längerfristige Speicherung von Befragten-IDs, deren Adressen vor Missbrauch zu schützen sind. Weitere datenschutzrechtliche Fragen und die Anforderungen an die Anonymisierung von Forschungsdaten werden in Kapitel 5.3 im Zusammenhang mit der Archivierung und Weitergabe sozialwissenschaftlicher Daten thematisiert.

1.4 Verantwortlichkeiten im Forschungsdatenmanagement

Verantwortung für das Datenmanagement trägt nicht nur der einzelne Primärforscher. Vielmehr sind mehrere Akteure in unterschiedlichen Rollen in das Forschungsvorhaben eingebunden. Aufgaben und Verantwortlichkeiten im Datenmanagement sollten deshalb frühzeitig, eindeutig und transparent in einem Projekt- und / oder Datenmanagementplan festgehalten werden (siehe folgende Übersicht). Dies gilt umso mehr, wenn in Kooperationsprojekten mehrere Forscher oder Forschergruppen sowie institutionelle oder externe Infrastruktureinrichtungen und Dienstleister zusammenarbeiten.

Übersicht 1: Aufgaben und Verantwortlichkeiten im Datenmanagement

- Forschungsdesign, Datenmanagementplan und Projektleitung
  Akteure: Primärforscher. In größeren Umfrageprogrammen ist es üblich, das Projekt durch spezielle Gremien zu leiten (wissenschaftlicher Beirat, Lenkungsausschuss o. ä.).
- Fragebogenentwicklung, Methodenfragen sowie Organisation und Dokumentation der Daten
  Akteure: Primärforscher und / oder spezialisierte Forscherteams.
- Einhaltung des Datenschutzes im Forschungsprozess
  Akteure: Primärforscher und Erhebungsinstitut.
- Erhebung, Erfassung, Aufbereitung, Kontrolle der Felddaten
  Akteure: Primärforscher und Erhebungsinstitut.
- Softwareeinsatz und Bereitstellung von Plattformen zum gesicherten Informations- und Datenaustausch im Projekt
  Akteure: Projektmitarbeiter und / oder institutionelle IT Dienste.
- Administrative und technische Datenorganisation, -sicherheit und -speicherung
  Akteure: Projektmitarbeiter und / oder institutionelle IT Dienste.
- Datensicherung und -bereitstellung nach Projektende
  Akteure: Projektleitung und institutionelle Dateneinrichtung bzw. Forschungsdatenzentrum oder Datenarchiv der Fachdisziplin.

1.5 Sach- und Personalkosten des Datenmanagements von Forschungsdaten

Die Kosten des Datenmanagements in einem Forschungsprojekt sind ein zentraler Faktor, der frühzeitig bei der Projektplanung berücksichtigt werden sollte. Angaben zu den Datenmanagementkosten fließen dann auch entsprechend in einen Drittmittelantrag ein.

Zu den datenbezogenen Personal- und Sachmitteln, die in den Datenmanagementplan aufgenommen werden sollten, zählen alle Aufwendungen von der Datenerhebung bis hin zu den Kosten einer qualitätsgesicherten Aufbereitung, Kontrolle und Dokumentation aller Materialien, die für die Vorbereitung, Durchführung und Veröffentlichung von Datenanalysen notwendig sind. Weiterhin sind neben den Aufwendungen für die projektbezogene Datensicherung auch eventuelle Kosten zur Vorbereitung der Daten und Metadaten für eine dauerhafte Sicherung in den Budgetplanungen zu berücksichtigten.

Dabei ist zu beachten, dass es in der akademischen Sozialforschung üblich ist, die Feldarbeit (großer) wissenschaftlicher Erhebungsprojekte kommerziellen Erhebungsinstituten zu übertragen. Die im Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. (ADM) organisierten privatwirtschaftlichen Unternehmen bieten ihre Dienstleistungen z. B. zur Datenerhebung auf der Grundlage professioneller Qualitätsstandards (ISO 20252:2006), ethischer Prinzipien (ESOMAR Kodex) und Richtlinien zur Durchführung unterschiedlicher Erhebungsverfahren an (vgl. ADM in Anhang A.2).

Gleichzeitig unterliegen kleinere Forschungsprojekte, die Daten in eigener Verantwortung erheben, gerade bei der Datenaufbereitung analogen Anforderungen an ein qualifiziertes Datenmanagement, um hochwertige Forschungsdaten zu produzieren.

In diesem Zusammenhang geben Schnell, Esser, Hill (2011: 481ff) Hinweise zu Kostenaspekten, die bei Ausschreibung einer standardisierten Datenerhebung durch ein Erhebungsinstitut zu beachten sind. Weiterhin beschreiben sie die wichtigsten Aspekte, die in einem Vertrag zur Durchführung der Datenerhebung und Erfassung geregelt werden sollten (z. B. Grundgesamtheit, Pretest, Art des Datensatzes, Dokumentation der Erhebung).

Eine grobe und nicht abschließende Aufteilung von möglicherweise auftretenden Kosten jenseits der üblichen Forschungstätigkeiten und -prozeduren vermittelt die folgende Übersicht.

Übersicht 2: Allgemeine Kostenaspekte (Sach- und Personalkosten) des Datenmanagements

- Projektmanagement: Verantwortlichkeiten zur Erstellung, Implementierung und Sicherung von Strategien und Verfahren des Forschungsdatenmanagements;
- Technische Sachmittel (Software; Hardware) und administrative Maßnahmen zur strukturierten und standardisierten Datenorganisation, -austausch, -speicherung und -sicherheit.
- Übersetzung von Fragebögen (Landessprache(n)) und / oder weiteren Studienmaterialien;
- Datenaufbereitung (Kontrolle, Bereinigung) und Dokumentation der Datenmodifikationen;
- Standardisierte Dokumentation auf Studien- und Variablenebene (Metadaten);
- Nutzung von technischen Dokumentationsstandards (z. B. DDI) und Werkzeugen;
- Anforderung an die Datenanonymisierung;
- Digitalisierung von Materialien zur Weiterverarbeitung bzw. Vorbereitung der nachhaltigen Sicherung der Studie und der erzeugten Daten (z. B. Messinstrument, Show-Cards).

Eine konkrete Kostenschätzung der zuvor genannten Aspekte des Datenmanagements hängt von weiteren Faktoren des konkreten Projektkontextes ab, die anhand folgender Leitthemen abgeschätzt und kalkuliert werden müssen.

Übersicht 3: Leitthemen zur Kostenabschätzung des Datenmanagements

- Grad des Erfahrungswissens der Projektmitarbeiter und der Integration von Maßnahmen und Verfahren zum Datenmanagement in bestehende Forschungsabläufe.
- Grad der Komplexität des Forschungsvorhabens (Untersuchungsdesign, Datentypen, Befragte);
- Grad der Komplexität der Projektorganisation (Mitarbeiter; Kooperationspartner);
- Definition der Anforderungen bzw. Aufgaben im Datenmanagement bzw. deren zentrale und / oder dezentrale Durchführung (je nach Projektorganisation);
- Anforderungen an technische und administrative Maßnahmen;
- Planung und Formen der Datendokumentation und Datensicherung;
- Bestehende Optionen zur nachhaltigen Datenarchivierung und –bereitstellung.

1.6 Sicherung, Archivierung und Weitergabe von Forschungsdaten

Eine vorausschauende Planung zum Umgang mit den Daten im Projektverlauf kann dazu beitragen, dass am Projektende nur noch notwendige Abschlussarbeiten zur Sicherung bzw. Weitergabe der Daten durchzuführen sind. Dies gilt umso mehr, wenn Abschlussberichte, Veröffentlichungen oder ein neuer Projektantrag bereits viel Zeit in Anspruch nehmen. So sollte so früh wie möglich entschieden werden, welche Daten, Dokumentationen und Projektergebnisse am Projektende gesichert werden müssen, um eine nachhaltige Nutzung der Daten zu ermöglichen. Dabei sind die bereits genannten Empfehlungen oder Förderrichtlinien der DFG zu berücksichtigen (vgl. Abschnitt 1.7). Die dort formulierten allgemeinen Vorgaben lassen sich durch konkrete Überlegungen zur Sicherung und Weitergabe von Forschungsdaten präzisieren, worauf in Kapitel 5 näher eingegangen wird. Auch in einer frühen Planungsphase können bereits entsprechende Beratungs- und Kooperationsmöglichkeiten im fachlichen Umfeld einbezogen werden, um effiziente Vorgehensweisen und Lösungen vorausschauend einzuplanen.

Die organisatorisch-technischen Möglichkeiten können sich von einer zehnjährigen Speicherung z. B. am Forschungsinstitut bis hin zur Langzeitarchivierung und Datenbreitstellung durch eine datenhaltende sozialwissenschaftliche Einrichtung erstrecken.

Bei der Suche nach einer organisatorischen Lösung sind auch Aspekte der zeitlichen Dauer der Datensicherung und Fragen der Datennutzung nach dem Ende des Forschungsvorhabens von wesentlicher Bedeutung. Intensivere Überlegungen zur Datenbereitstellung sind insbesondere dann notwendig, wenn die Daten langfristig in einem fachspezifischen Forschungsdatenzentrum oder im GESIS Datenarchiv für Sozialwissenschaften archiviert und für Sekundäranalysen zur Verfügung stehen sollen.

In diesem Zusammenhang sind auch rechtliche Aspekte des Datenschutzes hinsichtlich der Anforderungen an die Anonymisierung personenbezogener Daten zu klären. Weiterhin sind Fragen zur Nutzung und Verbreitung des archivierten Materials im Zusammenhang mit dem Urheberrechtsgesetz zu beachten.

Je nachdem, welcher zeitliche und organisatorische Rahmen der Datensicherung und Datenbereitstellung angestrebt wird, ist zu überlegen, ob bzw. wie das Forschungsprojekt die aufgeworfenen Fragen auch vertraglich mit der Datenserviceeinrichtung regeln kann. Aspekte des Rechtemanagements und des Datenzugangs werden anhand des Archivierungsvertrags des GESIS Datenarchivs thematisiert (vgl. Abschnitt 5.4).

1.7 Datenmanagement und Forschungsförderung der DFG

Nationale und internationale Förderinstitutionen erwarten immer häufiger, dass in Förderanträgen dargelegt wird, wie mit den Forschungsdaten in einem Projekt umgegangen werden soll und ob relevanten Daten nach Projektabschluss für eine Nachnutzung zur Verfügung stehen. Die Beschreibung entsprechender Maßnahmen wird dann in einem Datenmanagementplan (DCC 2010; NSF 2010) zusammengefasst. Auch die DFG fordert seit 2010 in ihren Antragsrichtlinien die Beschreibung von Maßnahmen zum Datenmanagement im Laufe des Forschungsprojektes:

- „Wenn aus Projektmitteln systematisch (Mess-)Daten erhoben werden, die für die Nachnutzung geeignet sind, legen Sie bitte dar, welche Maßnahmen ergriffen wurden bzw. während der Laufzeit des Projektes getroffen werden, um die Daten nachhaltig zu sichern und ggf. für eine erneute Nutzung bereit zu stellen. Bitte berücksichtigen Sie dabei auch – sofern vorhanden – die in Ihrer Fachdisziplin existierenden Standards und die Angebote bestehender Datenrepositorien." (zuletzt in DFG 2012c: 6)

Welche Daten bzw. Metadaten in einem Datenmanagementplan zu berücksichtigen sind, ist also jeweils im Rahmen des jeweiligen Projektziels und auf Grundlage der konkret genutzten bzw. erzeugten Forschungsdaten zu beantworten. Die Checkliste zur Datenmanagementplanung (S. 18) bietet einen Überblick über zentrale Aspekte und Leitfragen, die bei der Erstellung eines Datenmanagementplans Berücksichtigung finden können.

Weiterhin hat die DFG einen Leitfaden für die Beantragung von Langfristvorhaben herausgegeben, der auch Vorhaben in den Geistes- und Sozialwissenschaften behandelt. Im Sinne der Vorbereitung eines sozialwissenschaftlichen Langfristprojektes sollte demnach dargestellt werden,

- „welche praktischen Maßnahmen zur Pflege eines Panels und zum kontinuierlichen Datenmanagement vorgesehen sind. Dazu zählen auch Aussagen darüber, welche institutionellen Vorkehrungen getroffen werden, um eine Weiterführung der Studien unter anderen Personen als den Erstuntersuchern zu sichern. Die Planung der Datendokumentation muss die langfristige Nutzung der Daten ermöglichen." (DFG 2011a: 3)

Darüber hinaus unterstützt die DFG Sonderforschungsbereiche zusätzlich durch das Teilprojekt „Informationsinfrastruktur" (DFG 2012b: 8f). Damit soll das systematische Management der relevanten Daten und Informationen im Sonderforschungsbereich durch

- die Entwicklung und Umsetzung eines entsprechendes Konzeptes,
- die Bereitstellung einer leistungsfähigen Infrastruktur und
- die langfristige Nutzung der erschlossenen Daten nachhaltig und effektiv gefördert werden.

Ein Teilprojekt Informationsinfrastruktur kann sich dabei – einzeln oder gemeinsam – auf Ziele wie die Entwicklung von kollaborativen Arbeitsumgebungen und entsprechenden interoperablen Komponenten sowie auf die professionelle Speicherung, Erschließung, Bereitstellung im Sinne einer Langzeitarchivierung von Forschungsdaten über das Projektende hinaus beziehen (DFG 2012a).

1.8 Checkliste zum Forschungsdatenmanagement

Anforderungen und Fragestellungen des Datenmanagement	Kap.
1. **Drittmittelanträge & Förderrichtlinien**:	1.7
▪ Welche Anforderungen an das Datenmanagement sind zu berücksichtigen?	
2. **Datenbeschreibung:** Beschreibung von Eigenart und Umfang der Daten, die im Projekt erhoben und/oder aus externen Quellen stammen und bearbeitet werden:	1.1
▪ Welche Art von Daten werden wie erhoben (Erhebungsdesign, Befragte, Variablen, u. ä.)?	
▪ Werden vorhandene Daten in die Untersuchung einbezogen?	
▪ Zu welchem Zweck werden die Daten genutzt (Forschungsfrage, Hypothesentest; Re-Analysen, Replikation, Auftragsforschung, etc.)?	
▪ Welche besonderen Anforderungen stellen die Daten z. B. hinsichtlich der Erhebung, Weiterverarbeitung, Dokumentation, o. ä.?	
▪ In welchen technischen Formaten (Daten- bzw. Dateiformate) sollen die Daten erzeugt, unterhalten und ggf. verfügbar gemacht werden?	3.
3. **Metadaten und Dokumentation**	1.2
▪ Welche Arten von Metadaten werden im Projekt erzeugt und für die Nachnutzung der Daten dokumentiert, z .B. Methodendesign, Messinstrument, Feldarbeit, Datensatzstruktur, Syntaxdateien u. Ä.?	5.2
▪ Werden sozialwissenschaftliche Klassifikationen o. ä. bzw. technische Standards bei Erhebung u. Dokumentation eingesetzt?	4.
4. **Datenschutz**: Darstellung der Maßnahmen zum Datenschutz und ethische Fragen:	1.3
▪ Welche datenschutzrechtlichen Maßnahmen sind im Zusammenhang der Datenerhebung und Datenverarbeitung personenbezogener Daten zu berücksichtigen u. durchzuführen?	5.3
▪ Sind ethische Fragen besonders zu berücksichtigen?	
5. **Qualitätssicherung**: Beschreibung der Maßnahmen, Regeln und Prozesse, die zur Wahrung der Datenqualität im Projektverlauf eingesetzt und dokumentiert werden:	2
▪ Welche Verfahren und Regeln der Datendefinition, -kontrolle und -bereinigung werden eingesetzt und wie werden sie überprüfbar und nachvollziehbar dokumentiert?	3
Maßnahmen zur Organisation und zum technischen Schutz der Daten bzw. Dateien:	
▪ Sind (technische) Datenschutzmaßnahmen zu ergreifen? Wer darf auf Dateien zugreifen, sie verändern oder löschen?	3.1
▪ Wie werden die Dateien logisch organisiert (Konventionen, Versionskontrolle etc.)?	3.2
▪ Wie werden die Dateien gesichert und vor Verlust geschützt, z. B. durch Speicherverfahren und Backup der Dateien?	3.3

Anforderungen und Fragestellungen	Kap.
6. **Aufgaben & Verantwortung**: Datenmanagementaufgaben und verantwortlichen Personen: ▪ Welchen Umfang hat das Projekt und wie ist es organisiert? ▪ Welche Aufgaben sind im Datenmanagement zu bewältigen? ▪ Wie werden die Maßnahmen und Verantwortlichkeiten organisiert?	1.4
7. **Sach- und Personalkosten im Datenmanagement**: Erfassung der Sach- und Personalkosten für die Datenaufbereitung, Datendokumentation, Datensicherung im Projektverlauf sowie Vorbereitungen zur nachhaltigen Sicherung der Studie am Projektende. ▪ Sind in den Kosten Maßnahmen berücksichtigt, um die Datenqualität im Projektverlauf zu sichern und nachvollziehbar zu beschreiben?	1.5
▪ Sind Absprachen mit einer Serviceeinrichtung zur langfristigen Datennutzung geplant?	5.1
▪ Ist die Erstellung notwendiger Dokumentationen in der Budgetplanung berücksichtigt, um die Nachnutzung und intersubjektive Überprüfbarkeit der Daten zu gewährleisten?	5.2
8. **Datensicherung, Langzeitarchivierung und Datenbereitstellung**: Darstellung der Verfahren zur nachhaltigen Sicherung oder zur Langfristarchivierung und Bereitstellung der Studie (Daten und Dokumentationen):	1.6
Welche organisatorischen Optionen zur Sicherung der Studie am Projektende sind verfügbar und welche Lösung wird gewählt?	5
▪ Können datenhaltende Einrichtungen der Fachdisziplin, Regeln oder Beratungsmöglichkeiten zur Archivierung genutzt werden?	5.1
Sind die Daten und der Entstehungskontext für die langfristige Datennutzung hinreichend dokumentiert?	5.2
Welche rechtlichen Aspekte müssen bei der Sicherung oder Langfristarchivierung berücksichtigt und ggf. vertraglich geregelt werden? ▪ Sind besondere Maßnahmen zum Datenschutz vor der Archivierung bzw. im Zusammenhang mit der Datenbereitstellung zu planen? ▪ Sind urheberrechtliche oder patenrechtliche Rechte Dritter zu klären?	5.3
Regelung der Nutzungs- und Zugangsrechte im Zuge der langfristigen Archivierung und Bereitstellung der Daten: ▪ Unter welchen Bedingungen können die Daten bereitgestellt werden? ▪ Sind Embargos oder andere Einschränkungen zu berücksichtigen? ▪ Ist die Studie für einen bestimmten Nutzerkreis von Interesse?	5.4
Welche Daten und Dokumente sollen nach Projektende wie lange gesichert und für die weitere Nutzung bereitgestellt werden? ▪ Sind eventuell weitere Datenbestände zu sichern? ▪ Nach welchen Kriterien werden die Daten und Dokumentationen ausgewählt? ▪ In welchen Dateiformaten sollen Datenfiles und Materialien gesichert werden?	5.5

2 Datenaufbereitung und Datendokumentation

Voraussetzung für die computerbasierte statistische Analyse der erhobenen Daten ist die Erstellung eines analysefähigen Datensatzes. Er basiert auf den verschiedenen Forschungsphasen und Arbeitsschritten im Projektablauf:

- Fragebogenerstellung und Datendefinition mit Hilfe eines Codeplans,
- Pretests und entsprechende Anpassung von Fragebogen und Datendefinition,
- Durchführung der Datenerhebung (Feldarbeit),
- Erstellen des Rohdatenfiles mit den erhobenen Originaldaten (manuelle oder technische Datenerfassung bei klassischer PAPI Technik),
- Datenaufbereitung mit Datenkontrolle und -korrektur der erhobenen Originaldaten,
- Weitere Datenaufbereitungsschritte zum Aufbau des fertigen Analysedatensatzes.

Leitlinien und Konventionen zur Datendefinition und die Entwicklung der Datensatzstruktur stehen im Mittelpunkt des ersten Abschnitts dieses Kapitels. Der zweite Abschnitt widmet sich den Aspekten der Datenkontrolle und der Datenbereinigung der erhobenen Daten. Diese vier Aufgaben stellen Kernanforderungen an das unmittelbare Management der erhobenen Daten dar, die im Zuge einer systematischen Datenaufbereitung und Datendokumentation erfüllt werden müssen.

Die Nutzung möglichst einheitlich strukturierter Regeln und Verfahren sollte bereits frühzeitig mit der der Datendefinition beginnen. Die Ergebnisse systematischer Vorgaben zur Datenbeschreibung können dann auch (zeitschonend) während der späteren Datenkontrolle und der Datenbereinigung genutzt werden. Gut geplante Regeln von der Datendefinition bis hin zur Datenbereinigung sichern damit auch das Projektziel, qualitätsgesicherte und nachvollziehbar dokumentierte Daten für die Datenanalyse zu generieren. Zugleich sind damit wichtige Voraussetzungen geschaffen, um die Forschungsdaten auch über das Projekt hinaus nachhaltig für weitere Forschungszwecke nutzen zu können.

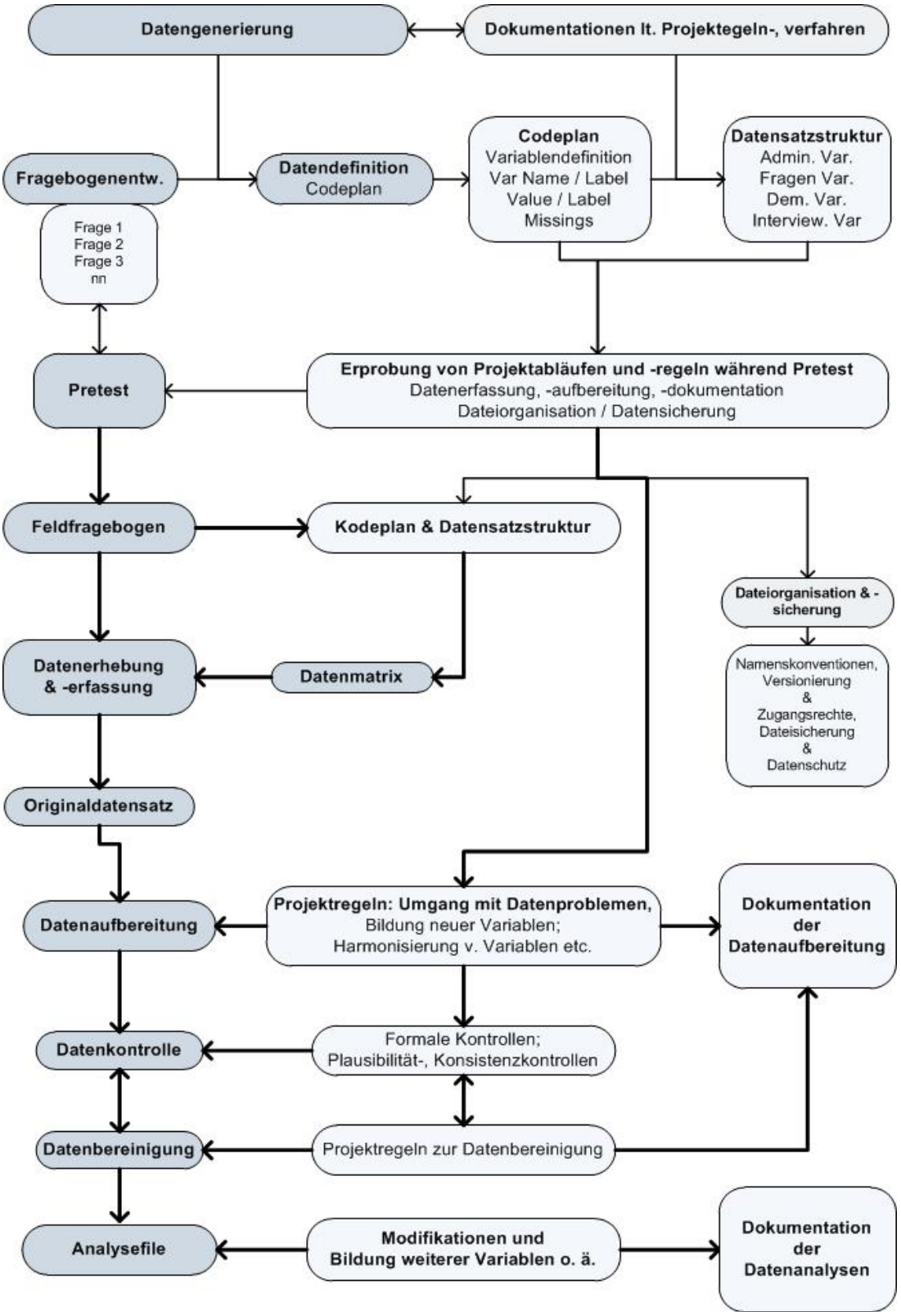
Entsprechende Maßnahmen und Methoden werden im Folgenden anhand von Leitlinien aus der Praxis des GESIS Datenarchivs aufgegriffen. Die in Arbeitspapieren und Projektdokumentationen beschrieben Vorgehensweisen wurden in vielen Fällen im Rahmen von Kooperationsprojekten mit Forschern und Forschungsnetzwerken gemeinsam entwickelt und jeweils projektspezifisch umgesetzt. In ihren allgemeinen Grundsätzen entsprechen diese Leitlinien zugleich Best Practice Empfehlungen sozialwissenschaftlicher Datenarchive anderer Länder (u. a. ICPSR 2012, UKDA 2011).

Der Fokus der folgenden Abschnitte liegt dabei auf grundlegenden Aspekten der Aufbereitung und Dokumentation einfacher Datensätze. Spezifische Leitfäden und Regeln verwiesen beispielhaft auf das Datenmanagement komplexer Datensätze wie dem ALLBUS oder auf entsprechende Standards bei der Aufbereitung international vergleichender Studien.

Die konkrete Planung aller Detailschritte der Datenaufbereitung und Datendokumentation ist immer von der Spezifik der Daten, den im Projekt festgelegten Regeln des Datenmanagements, der eingesetzten Software und den verfügbaren Ressourcen abhängig. Die folgenden Informationen sind deshalb als allgemeine Empfehlungen zu verstehen.

Die Abbildung 2 beschreibt zunächst wesentliche Zusammenhänge – von der Datendefinition bis hin zur Erstellung des Analysefiles – in einer Übersicht. Eingebettet in den zugrundliegenden Forschungsplan bildet die Organisation der Abläufe und die Festlegung entsprechender Richtlinien zur Definition, Erhebung, Aufbereitung und Dokumentation der Daten den Kern des projektspezifischen Forschungsdatenmanagements.

Abbildung 2: Vom Fragebogen zum Analysefile



2.1 Der Fragebogen: Grundlage der Datendefinition und der Datensatzstruktur

Das Vorgehen in der Dokumentation und Aufbereitung von Forschungsdaten ist stark von der gewählten Erhebungstechnik (PAPI: Paper & Pencil Interviewing; CAI: Computer Assisted Interviewing) abhängig. Überwiegend werden heute bei der Datenerhebung computergestützte Verfahren eingesetzt. CAPI und CATI Software wird u. a. auch eingesetzt, um den Ablauf leichter zu steuern (z. B. Filterführung), Antworten unmittelbar elektronisch zu erfassen, den Aufwand für die Datenkontrolle zu verringern und dadurch eine schnelle Weiterverarbeitung der Daten zu ermöglichen. Zusätzlich werden durch dieses Vorgehen auch die Fehlerquellen der manuellen Datenerfassung ausgeschlossen. Anderseits stehen diesen Vorteilen zur Sicherung der Datenqualität u. a. die Kosten gegenüber, die für Entwicklung, exakte Programmierung, Tests von Ablauf- und Kontrollroutinen sowie für die Erstellung der erforderlichen Dateneingabemasken erforderlich sind (vgl. Lück, Baur 2011: 28ff).

Liegen die Angaben aus der Erhebung in Form ausgefüllter Fragebögen vor, müssen die Befragungsergebnisse im Zuge der Datenerfassung zunächst in einen maschinenlesbaren Datenfile übertragen werden. Voraussetzung für die Erfassung und Aufbereitung dieser Angaben in Form eines ersten Rohdatensatzes ist die Erstellung eines Codeplans und einer Datenmatrix.

Wurde die Datenerhebung computergestützt durchgeführt, liegen die Daten bereits in Form eines strukturierten Datenfiles vor und können mit einem Statistikprogramm weiter aufbereitet werden. Aber auch in diesem Fall müssen die Daten zur Vorbereitung der Programmierung zuerst definiert sowie die jeweiligen Filterbedingungen und Ablaufstrukturen festgelegt werden. Außerdem ist nicht zu vergessen, dass - wie im Fall einer PAPI basierten Datenerhebung - auch der mit einem CAI-System erstellte Rohdatenfile systematisch während der Datenaufbereitung kontrolliert werden muss.

Die Erstellung des Rohdatenfiles kann als die strukturierte Abfolge von drei Hauptschritten beschrieben werden, die auf dem Master- oder Feldfragebogen aufbauen (vgl. 2.1.2-2.1.4). Mit der Entscheidung über die endgültige Fassung des Fragebogens werden Struktur und Inhalte für die Variable(n) vorgegeben, die auf einer Frage beruhen.

Deshalb ist es hinsichtlich der Planung der Datenaufbereitung sinnvoll, parallel zum (Master-) Fragebogen so früh und vollständig wie möglich alle Variablen und ihre Bestandteile zu definieren und in einem Codeplan zu dokumentieren. Gleichzeitig können auch zusätzlich zu bildende Variablen (u. a. administrative Variablen) frühzeitig aufgenommen werden. Die wesentlichen Arbeitsschritte und Regeln zur Erstellung von Codeplan und Datenmatrix sowie Formen der Datenerfassung werden im Folgenden beschrieben.

2.1.1 Erstellen des Codeplans - von der Fragenstruktur zur Datensatzstruktur

Der Codeplan ist eine strukturierte und klartextbasierte Liste, mit der die zu erhebenden Daten zunächst eindeutig als Variablen definiert werden, um den Aufbau des Datensatzes für die weitere maschinelle Verarbeitung der Daten vorzubereiten.

Dabei kann die Bildung von verschiedenen Variablenarten (A) unterschieden werden, deren Abfolge (B) abschließend festgelegt wird. Die im Folgenden dargestellten Variablen stellen ein (nicht abschließendes) Spektrum von Variablenarten und Typen dar, die je nach Komplexität der Studie und des zugrundeliegenden Fragebogens zum Einsatz kommen können.

A. Definition und Beschreibung unterschiedlicher Variablenarten

1. Variablen zur Erfassung und Verwaltung der erhobenen Daten

Der Codeplan enthält am Anfang einen Platzhalter (ID-Variable) zur Codierung der Identifikationsnummer für jeden Fragebogen (Fall, Befragter). Die ID muss eindeutig sein, um partielle Korrekturen, Ergänzungen, Re-Interviews (z. B. Panel) etc. im Projekt durchführen zu können.

Im Codeplan können je nach Bedarf zusätzliche Variablen definiert werden, die der Verwaltung des Datenfiles dienen, z. B. eine

- Projekt-ID Variable; unerlässlich, wenn mehrere Erhebungen verwaltet werden sollen,
- Datenfile-ID Variable für Daten je Personengruppe, Welle und/oder Länderdatensatz,
- Versions-ID Variable zur eindeutigen Kennzeichnung der Version des Datenfiles,
- Befragte ID Variable zur Fallidentifikation in einem Datenfile,
- Fragebogen ID Variable zur Kennzeichnung von Fragebogensplits oder Sprachversionen,
- Gewichtungsvariablen.

2. Variablen zu jeder Frage aus dem Fragebogen:

2.1. Fragen (-Modul) -Variablen

Generell wird jeder Frage bzw. jedem Frageitem genau eine Variable im Datensatz zuordnet. Die Variable wird durch den Variablennamen (Variable Name) und das Variablenetikett (Variable Label) beschrieben. Die Zuweisung von Variablen zu Fragen folgt idealerweise dem Ablauf der Fragen im Fragebogen.

- Jeder Antwortkategorie einer Frage wird ein eindeutiger numerischer Code (Value) in der entsprechenden Variable zugewiesen. Die Bedeutung des Codes wird durch eine Werteetikette (Value Label) mit Bezug auf den Inhalt der Antwortkategorie dokumentiert.
- Zusätzlich wird jede Art fehlender Werte kategorisiert, indem ihnen ein bestimmter Code (Missing Value) und eine Bedeutung (Missing Value Label) zugeordnet werden. D. h., fehlende Werte sollten vollständig als solche definiert werden und Auskunft über ihre (unterschiedliche) Herkunft geben.
- Eine spezielle Kategorie fehlender Werte ergibt sich aus Filterfragen, für die entsprechende „trifft nicht zu" Kategorien im Codeplan für die betroffenen Folgevariablen vorgesehen werden müssen.
- Schließlich wird mit der / den Spaltennummer(n) angegeben, wo der numerische Wert für die jeweilige Antwort in der später noch zu erstellenden Datenmatrix eingetragen bzw. im fertigen Datenfile aufgefunden werden soll.

Besonderheiten bei der Variablenbildung ergeben sich, wenn in Abhängigkeit von der Art der Fragekonstruktion eine entsprechende Art und Anzahl von Variablen gebildet werden muss.

- Geschlossene Frage mit Einfachnennung: Um eine Frage mit einfacher Antwortmöglichkeit zu codieren, wird jeder Frage eine Variable zugeordnet und die Antwortalternative mit einem numerischen Wert (ja=1; nein =2) codiert.

  Mehrstufige Antwortskalen werden entsprechend der Anzahl von Antwortvorgaben codiert.

- Geschlossene Frage mit Mehrfachnennungen: Bei mehreren Unterfragen, Statements, Fragelisten wird für jede einzelne Fragestellung / Antwortkategorie eine „Dummy-Variable" mit dichotomer Merkmalsausprägung (z. B. Trifft zu / trifft nicht zu) angelegt.
- Jede einzelne Frage (Statement o. Ä.) in einer Itembatterie wird in einer Variablen abgebildet und gemäß den Antwortmöglichkeiten (z. B. einer Skala) codiert.
- Offene Fragen mit Textangaben („Weshalb haben Sie Ihren Arbeitsplatz verloren?") müssen zunächst nach der Feldarbeit gesondert ausgewertet werden. Dazu werden zuerst alle unterschiedlichen Antworten erfasst und dann nach einem Schema kategorisiert.

  Standardisierte Codeschemata stehen z. B. für die Codierung von Angaben zum Beruf mit dem ISCO (International Classification of Occupation) zur Verfügung.

2.2. Demographische Variablen

- Offene Fragen mit numerischen Angaben (z. B. Einkommen in Euro) werden als numerischer Wert übernommen und die Maßeinheit einmalig im Variable Label dokumentiert.

Die Zusammenfassung (Kategorisierung) von Informationen (z. B. Alter, Einkommen) sollte über zusätzlich gebildeten Variablen mit höherem Aggregationsniveau erfolgen. Die zugrundeliegende Originalvariable (oder weitere bzw. andere Kategorisierungen in den Rohdaten) sollte im Interesse der langfristigen Nutzbarkeit der Daten erhalten bleiben.

Im Kontext international vergleichender Studien stellt z. B. das ISSP (International Social Survey Programme) seinen Mitgliedern standardisierte Definitionen für die sog. „Hintergrundvariablen" und modulbezogene Codierungsschemata zur Verfügung (ISSP 2010b).

2.3 Erhebungs- und Interviewer-Variablen

- Interviewer-ID und Interviewprotokoll-Variablen (Interviewdatum, -beginn, -ende, etc.).

2.4 Bildung und Beschreibung zusätzlicher Variablen

Zusätzlich konstruierte Variablen (z. B. Aggregatvariablen zum Einkommen) oder harmonisierte Variablen sollten im Anschluss an die jeweilige Ursprungs- oder Originalvariable einfügt werden. Gleiches gilt für abgeleitete Variablen wie z.B. der Inglehart-Index oder Haushalts- und Familientypologien, die gegebenenfalls in der Abfolge der Ursprungs-Variablen eingefügt werden können (vgl. z. B. ALLBUS 2008a - Variable Report: Xi und 105).

Je nach Umfragedesign, Typ der Studie oder Typ der Frage können Variablen zusätzlich (formal oder inhaltlich) typisiert gebildet werden, z. B. durch die einheitliche Kennzeichnung von wiederholt erhobenen Fragen mit einheitlichen Variablen Namen und / oder Labeln für die entsprechenden Trendvariablen.

B. Festlegung der Reihenfolge der Variablen im Datensatz

Um einen Datensatz übersichtlich zu gestalten, hat es sich in der Praxis als sinnvoll erwiesen, Variablen nach formalen Kriterien in Gruppen einzuteilen, wie sie in den vorherigen Abschnitten bereits benutzt wurden. Die konkret aus dem Fragebogen abgeleitete Variablenstruktur ergibt sich dabei aus der Reihenfolge der Fragen bzw. den jeweiligen Fragentypen (Einzelfrage; Fragebatterie) und den zusätzlich gebildeten (technischen, harmonisierten, etc.) Variablen. Das folgende Beispiel veranschaulicht die mögliche Anordnung formal definierter Variablen bzw. -typen in einem Datenfile:

- Administrative und technische Variablen
- Variablen mit Bezug auf die Fragen im Fragenbogen (inhaltliche Variablen)
    - Fragen (-modul) -Variablen
    - Demographie Variablen
- Erhebungs- und Interviewer-Variablen

Die (frühzeitige) Festlegung von Gruppen zusammenhängender Variablen bzw. der Reihenfolge aller (geplanten) Variablen erleichtert zunächst die Datenerfassung. Eine konstante und übersichtlich organisierte Variablenstruktur verringert aber auch den Aufwand bei der Eingabe von statistischen Prozeduren während der Aufbereitung und Auswertung zusammengehöriger Variablen

2.1.2 Definition der Datenstruktur des Rohdatenfiles durch die Datenmatrix

Um die Erfassung der Originaldaten in einem Rohdatenfile vorzubereiten, muss auf Grundlage der Definitionen des Codeplans zunächst mit Hilfe eines entsprechenden Softwareprogramms eine Datenstruktur erstellt werden, die dann mit den erhobenen Daten gefüllt wird. Dies geschieht in der Regel in Form einer rechteckigen Datenmatrix. Darunter versteht man eine fix strukturierte Tabellenstruktur, die je Spalte eine Frage als Variable (= Merkmal) repräsentiert und eine Zeile alle Angaben eines Befragten (= Fall) zu den Fragen enthält. Die Antwort zu einer Frage wird durch numerische oder alphanumerische Werte in den entsprechenden Spalten abgebildet.

Die programmtechnische Erstellung einer Datenmatrix wird wesentlich durch die Art der Datenerfassung festgelegt, die im nächsten Punkt beschrieben wird.

2.1.3 Erfassung der erhobenen Daten in einem Rohdatensatz

Beim Einsatz computergestützter Techniken fallen Erhebung und Erfassung der Daten in einem Prozess zusammen. Bei der Nutzung von PAPI Erhebungstechniken kann zwischen der manuellen und der automatisierten Datenerfassung unterschieden werden.

Die Erstellung der Datenmatrix und die manuelle Dateneingabe können mit verschiedenen Programmen vorgenommen werden:

- Komplexe Datenverwaltungssoftware mit Möglichkeiten zur Programmierung von Datenerfassungsmasken, z. B. dem SPSS Modul Data Entry, EpiData, Access, CSPro o. ä.
- Dateneditoren in Statistikprogramme mit (kombinierten) Funktionen zur Datendefinition und Dateneingabe. So bietet der SPSS Dateneditor eine Variablenansicht zur formalen Datendefinition und die Datenansicht zur Dateneingabe und Datendarstellung. Diese Funktionen sind in ähnlicher Weise auch in anderen gängigen Statistikprogrammen wie SAS, STATA oder im Open-Source Programm R verfügbar.
- Datendefinition und -erfassung mit einem tabellenorientieren Programm wie z. B. Excel
- Datenerfassung mit einem einfachen Texteditor zur Erzeugung einer rechteckigen Rohdatenmatrix im ASCII-Format. Jede erfasste Angabe eines Falls stellt eine Zeile mit aufeinanderfolgenden Zahlen oder Buchstaben dar. Diese Art der Datenerfassung ist sehr fehleranfällig und wird nur noch in seltenen Fällen eingesetzt.

  Allerdings ist es üblich, dass Datensätze im ASCII Format von Datenserviceeinrichtungen für Sekundäranalysen zur Verfügung gestellt werden. Liegen die Rohdaten als ASCII Datei vor,

sind die zeilenweisen Abfolgen von Ziffern, Buchstaben und eventuelle Leerstellen zunächst ‚sinnlos'. Deshalb muss ein Codeplan bereitgestellt werden, um die Inhalte in Zeilen und Spalten zu verstehen und mit einem Statistikprogramm zu verarbeiten.

Auf Basis des Codeplans wird mit Hilfe einer softwarespezifischen Steuerdatei eine Datenmatrix definiert. Die Steuerdatei (Setup) wird gemeinsam mit dem Rohdatenfile in das Statistikprogramm eingelesen. Sie erlaubt die inhaltliche Zuordnung und strukturierte Verarbeitung der Informationen aus der ASCII Datei.

- Datenerfassung durch Belegleser: Die Datenerfassung kann auch automatisiert durch Belegleser und entsprechender Scannersoftware erfolgen (z. B. TeleForm; DOCUMENTS for Forms). Dazu müssen Anforderungen an die Fragebogengestaltung und die Programmierung der Software vorab berücksichtigt werden, um ‚zu erkennen', welche Variablen vorliegen und wo im Fragebogen eine Antwort angekreuzt wird.

2.1.4 Konventionen zur Festlegung von Variable Name und Variable Label

Bei der Festlegung von Variablennamen bzw. Gruppen von Variablen (z. B. technische, inhaltliche oder demographische) sollte berücksichtigt werden, dass sie das Hauptobjekt von Datenanalysen darstellen und deshalb in allen Bearbeitungsphasen im Projekt möglichst einfach, verständlich und eindeutig erkennbar sein sollten.

Entsprechend klare Konventionen vereinfachen zugleich die Erschließung oder allgemein die Nachnutzung von Daten. Zur Bildung der Variablennamen (A) lassen sich die im Folgenden beschriebenen Strategien anwenden, die aber immer gemeinsam mit erläuternden Variablen Labels (B) genutzt werden sollten. Das Vorgehen folgt der Praxis im GESIS Datenarchiv (Brislinger et al. 2009).

A. Festlegung von Variablennamen

1. Variablennamen mit Fragenummer

Bei diesem Vorgehen wird die Variable nach der Fragenummer benannt, z. B. F1 bis Fn. Dadurch wird ein direkter Bezug der Variable zur Originalfrage und deren Reihenfolge im Fragebogen abgebildet. Fragen, die die Bildung mehrerer Variablen erfordern, können entsprechend der Abfolge der abgefragten Kategorien oder Items benannt werden (F2.1, F2.2 ... oder F3a, F3b ...), auch wenn diese erweiterte Kennzeichnung z. B. bei Itembatterien nicht im Fragebogen vorgegeben ist.

2. Variablennamen mit aufsteigender Nummerierung

Eine übliche Namensgebung ist die aufsteigende Nummerierung mit dem leitenden Buchstaben V (Variable), z. B. V01 bis Vnn (Datensatz mit max. 99 Variablen), V001 bis Vnnn (max. 999 Variablen) etc. Damit wird eine einfache Reihenfolge der Variablen im Datensatz abgebildet. Diese Form des Variablennamens ist linear eindeutig. Die Variablen können aber nicht nach Inhalt bzw. Typ unterschieden werden (Fragenmodul, Demographiemodul, abgeleitete Indizes z. B. Einkommen).

Rein numerisch orientierte Namenskonventionen haben den Nachteil, dass sie keinen Bezug zu den Inhalten einer Variablen herstellen und so den kognitiven Umgang mit ihnen erschweren. Um dies zu vermeiden, werden oftmals inhaltlich festgelegte Kürzel als Gedächtnisstütze (Mnemonik) genutzt, damit Variableninhalte einfacher zu erkennen sind.

3. Mnemotechnische Variablennamen

Diese Kunstnamen, die den wesentlichen Inhalt der Variablen wiedergegeben (sollen), können eine einfache Merkhilfe darstellen. Ein solches Vorgehen bietet sich bei Längsschnittanalysen an, wenn Frage-Module wiederholt eingesetzt werden und die Variablen trotz unterschiedlicher Position im Fragebogen die gleichen Namen behalten sollen, z. B. "B_EKOM" für das Einkommen des Befragten.

Zu beachten ist, dass mnemotechnische Bezeichnungen für Dritte (innerhalb und außerhalb des Forschungsprojektes) nicht immer verständlich und bei großen Variablenzahlen schwer handhabbar sind. Bei mnemotechnischen Variablennamen könnte die Position der Frage im Fragebogen (Fragenummer) auch als zusätzliche Orientierungshilfe in das Variable Label aufgenommen werden.

4. Variablennamen mit Präfix, Stamm und Suffix

Bei der Benennung von Variablen komplexer Datensätze, etwa aus Länder vergleichenden Umfragen, werden zur Kennzeichnung thematischer oder strukturell unterschiedlicher Variablenblöcke (z. B. demographische Variablen) häufig Kombinationen aus Präfix, Stamm und Suffix verwendet.

Zahlreiche Beispiele liefern etwa die Standards für Hintergrundvariablen des International Social Survey Programme, z. B. die länderspezifische Variable zur Parteienneigung in Österreich „AT_PRTY" mit dem erklärenden Label „Country specific party affiliation: Austria" (ISSP 2010a).

5. Kombination von Elementen bei der Namensbildung von Variablen

Bei hoher Komplexität der Umfrage und einer Vielzahl von Variablen können auch verschiedene Elemente zur Namensbildung genutzt werden, u. a. um Besonderheiten für einzelne Länder kenntlich zu machen.

Das folgende konstruierte Beispiel zeigt eine Variable mit fünf Antwortkategorien, die nur in drei Ländern eingesetzt wurde. In allen anderen Ländern wurden vier Antwortkategorien benutzt. Zur Erläuterung wird das Variablenlabel benutzt, um den Bezug zur Frage und zu den abweichenden Antwortkategorien herzustellen:

- V154_4C "working mother warm relationship with children (Q46A) (4 answer category),
- V154_5C "working mother warm relationship with children (Q46A) (5 answer categories)".

Eine komplexere Art der Vergabe von Variablennamen entsteht, wenn diese systematisch mit thematischen Kategorien verknüpft werden und dann über mehrere Wellen eingesetzt werden, wie es z. B. der „EVS 1981-2008 Variable Report Longitudinal Data File" (EVS 2011: 29) umsetzt.

Bei der Festlegung der Variablennamen sollten auch später erforderlich werdende Erweiterungen bedacht werden. Nachfolgende Erhebungswellen, die Befragung von Subpopulationen oder abweichende Variablen in einzelnen Ländern können Ergänzungen des Namens der Ursprungs-Variablen erforderlich machen.

B. Festlegung von Variable Labels

Ziel der Variablenetiketten (Variable Labels) ist Variablennamen durch eine möglichst aussagekräftige und kurze Beschreibung des Inhalts der Variablen zu ergänzen. Zusätzliche Informationen zum Kontext oder zur Unterscheidung der Variablen, die in das Variable Label aufgenommen werden könnten, sind z. B.

- die Fragennummer aus dem Fragebogen;
- die Anzahl der Antwortkategorien oder
- Hinweise zu Art oder Besonderheiten der Variable: z. B. ob sie neu gebildet oder recodiert wurde, besondere Filter beinhaltet, die Variable länder- oder wellenspezifisch ist oder etwa Abweichungen vom Standard-Codeschema aufweist.

2.1.5 Codierung gültiger Werte der Antwortkategorien: Values und Value Labels

Die Variablen sollten in der Reihenfolge der dazugehörigen Fragen im Fragebogen vercodet werden, um den Codierungsablauf zu strukturieren und Kontrollen zu erleichtern. Variablen, deren Bearbeitung eine besonders hohe Konzentration sowie Hintergrundkenntnisse erfordern, wie etwa Berufscodierungen, sollte soweit möglich von einem erfahrenen Mitglied des Projektteams durchgeführt werden.

Aus der Perspektive einer langfristigen Nutzung der Daten ist es weiterhin empfehlenswert, national oder international akzeptierte Klassifikationssysteme zur Codierung von Ländern, Berufen, Bildungsangaben usw. einzusetzen. Gängige sozialwissenschaftlich relevante Klassifikationen und Standards werden im Abschnitt 4.1 vorgestellt. Die folgenden Codier- und Dokumentationsregeln entsprechen üblichen Konventionen, wie sie auch im GESIS Datenarchiv angewendet werden.

Konventionen zur numerischen Codierung von Antwortkategorien

Bei der Codierung werden numerische oder alpha-numerische Codes verwendet. Die Codes müssen die Ereignismenge der Frage im Fragebogen vollständig repräsentieren, sich gegenseitig ausschließen und eindeutig definiert sein. Gleichartige Ereignismengen werden entsprechend eines einheitlich festgelegten Codierungsstandards auch einheitlich vercodet.

Die Zuweisung der Werte zu den Antwortkategorien wird aufsteigend - mit 1 beginnend - in der Folge vorgenommen, in der die Antwortmöglichkeiten im Fragebogen vorgegeben sind. Hat der Code „0" die inhaltliche Bedeutung von „keine"' (z. B. Anzahl der Kinder), bleibt er erhalten.

Bei **Fragen mit Mehrfachnennungen** werden die einzelnen Antwortkategorien als dichotome Variable mit „0" und „1" codiert:

Beispiel:
Value 1 „genannt" oder „vorhanden"; Value 0 „nicht genannt" oder „nicht vorhanden".

Die Information „**Keine Antwort auf die Antwortkategorien der Frage**" kann für jede Variable in einem zusätzlichen Code abgelegt bzw. in einer neuen Variable („keine der **Antwortkategorien** genannt") abgebildet werden. Diese neu angelegte Kategorie / Variable ist von bereits im Fragebogen vorgegebenen vergleichbaren Kategorien zu unterscheiden.

Die Antworten auf **offene Fragen** sollten im Analysedatensatz kategorisiert vercodet sein. Ein entsprechendes Codeschema wird vorher wird in einem gesonderten Arbeitsschritt erstellt und dokumentiert. Zusätzlich sollten die Originalantworten zur Verfügung gestellt werden, z. B. als Word, Excel oder SPSS Datei.

Bei **Skalen** (Skalometer, Thermometer) werden Ausprägungen gegebenenfalls so umcodiert, dass dem niedrigsten negativen Wert die kleinste, dem höchsten positiven Wert die größte vorgesehene Zahl zugewiesen wird. Die Information über die ursprünglichen Kategorien wird im Value Label dokumentiert. Werden Items von -5 bis + 5 nach 1 bis 11 recodiert, sollte dies wie folgt dokumentiert werden.

Beispiel: Value 1 „(-5) lehne vollständig ab" / Value 11 „(+5) stimme vollständig zu".

Dokumentation der Antwortkategorien durch Value Labels

Eine möglichst exakte Übernahme des Textes der Antwortkategorien einer Frage als Value Label erleichtert die Arbeit mit den Daten im Projekt und spätere Sekundäranalysen durch andere Forscher. Auch formale Kontrollen von Datendefinitionen sollten mit Hilfe von Outputdateien durchgeführt werden, in denen die Variablenwerte mit Labels versehenen sind. Gleichzeitig sind solche aussagekräftigen Unterlagen bei Datenanalysen wesentlich leichter zu lesen, als die Bedeutung von Werten im Codeplan oder Fragebogen immer wieder nachzuschlagen. Dies gilt umso mehr, wenn Ergebnisse präsentiert, in Berichten dargestellt oder die Daten später zur weiteren Nutzung bereitgestellt werden sollen.

Jedoch sollten die Value Labels in Tabellen nicht zu lang sein (20 bis 40 Zeichen). Bei erforderlichen Kürzungen von langen Antwort-Texten aus dem Fragebogen ist darauf zu achten, dass Aussagekraft und Trennschärfe der Antwortvorgaben nicht verlorengehen.

2.1.6 Codierung fehlender Werte: Missing Values und ihre Value Labels

Fehlende Werte (Non-Responses) werden systematisch in Item-non-Response und Unit-non-Response unterteilt. **Unit-non-Response** sind vollständige Ausfälle von Befragungseinheiten (Fälle), wenn die Befragung einer Zielperson aus der Stichprobe nicht durchgeführt werden konnte, weil eine Teilnahme verweigert wurde oder die Person verstorben, nicht erreichbar oder wegen anderer Ursachen nicht befragbar war. Hierbei sind auch Projektkriterien festzuglegen, die anzeigen, ab welchem Ausmaß unvollständig ausgefüllte Fragebögen in diese Non-Response Kategorie fallen. Diese Informationen fließen als Angaben zur Ausschöpfungsquote in den Methoden- oder Feldbericht ein (vgl. dazu AAPOR 2008.).

Item-non-Responses beschreiben demgegenüber fehlende Werte, die aufgrund von Missverständnissen im Interview, fehlenden oder verweigerten Antworten usw. entstehen können. In diesem Zusammenhang sind auch die strukturell bedingten Antwortausfälle zu berücksichtigen, z. B. wenn Fragen einzelnen Befragtengruppen bzw. bei komplexen Studien in einzelnen (Teil-) Erhebungen, Ländern oder Wellen nicht vorgelegt wurden (s. u.). Die Deklaration und Codierung dieser fehlenden Werte wird im Folgenden näher erläutert.

Deklaration von Missing Values

Den fehlenden Werten werden in den jeweiligen Variablen spezielle Codes zugewiesen, die als „Missing Values" definiert sind oder werden. Ihre jeweilige Bedeutung wird durch passende Value Labels dokumentiert. Sie können, müssen aber nicht im Fragebogen vorgegeben sein. Teilweise basieren sie auch auf schriftlichen Intervieweranmerkungen aus der Befragung, die dann bei der Dateneingabe bzw. bei Datenkontrollen zusätzlich berücksichtigt werden können.

Alle Missing Values müssen im Zuge der Datendefinition vollständig deklariert werden, um eine strukturierte Datenkontrolle während der verschiedenen Arbeitsschritte in der Datenaufbereitung zu erleichtern. Die verschiedenen Item- oder Filter-non-Responses sollten - auch im Interesse der späteren Datenanalysen - möglichst differenziert erfasst und dokumentiert werden. Durch dieses Vorgehen können leere Zellen, also System Missings, ausgeschlossen werden.

Die Vermeidung von System Missings betrifft im Besonderen integrierte Datensätze, Trendreihen oder andere komplexe Daten. Hier sollten z. B. die erhebungsbedingten Gründe für das vollständige Fehlen der Antworten explizit benannt werden (z. B. das Auslassen von Fragen in einzelnen Ländern oder zu einzelnen Zeitpunkten), um einen transparent dokumentierten Datensatz zu erhalten.

Allgemeine Codierungsregeln von Missing Values

Für die Behandlung der fehlenden Werte innerhalb des Datensatzes werden entweder

- die höchsten numerischen Codes verwendet, die sich außerhalb des jeweiligen gültigen Wertebereiches der Variable befinden (Tabelle 2: Beispiel A und B) oder die
- fehlenden Angaben werden mit negativen Werten (Tabelle 2: Beispiel C) codiert.

Der Vorteil der Codierung mit negativen Werten liegt in einer visuell klarer wahrnehmbaren Abgrenzung von den positiven gültigen Werten. Dies vereinfacht etwa das Arbeiten mit Kontrollunterlagen oder Anpassungen von Steuerdateien (Syntaxdateien) und unterstützt Präsentationen von Ergebnissen.

Den einzelnen Missing Kategorien sollten für alle Variablen einer Studie oder Studienkollektion in Abhängigkeit von ihrem Wertebereich einheitliche Codes zugewiesen werden.

Übersicht 4: Codierungsbeispiele fehlender Werte

Beispiel A	Beispiel B	Beispiel C
Ist der Wert „7" Teil des gültigen Wertebereiches der Variable, wird „97" codiert. Ist „97" Teil des gültigen Wertebereiches, wird „997" codiert.	Hat der Code „0" eine inhaltliche Bedeutung oder werden dichotome Variablen mit „0" und „1" codiert, so wird „trifft nicht zu" auf den Code „9" gesetzt.	Codierung mit negativen Werten und englischsprachigen Labels
7 (bzw. 97, 997) verweigert	6 (bzw. 96, 996) verweigert	-1 don't know
8 (bzw. 98, 998) weiß nicht	7 (bzw. 97, 997) weiß nicht	-2 no answer
9 (bzw. 99, 999) keine Angabe	8 (bzw. 98, 998) keine Angabe	-3 not applicable
0 trifft nicht zu	9 (bzw. 99, 999) trifft nicht zu	-4 not asked in survey

Die Kategorie „keine Angabe" („not applicable") wird für Fälle verwendet, in denen

- keine Antwortkategorie oder fälschlich mehrere Kategorien markiert wurden,
- für Values, die sich außerhalb des definierten Wertebereiches einer Variable befinden,
- sowie bei Angaben, die vermutlich falsch sind und nicht rekonstruiert werden können.

Werden verschiedene Kategorien wie z. B. „keine Angabe" und „verweigert" zusammengefasst, sollte dies im Value Label dokumentiert werden.

Die Kategorien „andere" oder „sonstige" können als inhaltlich bedeutsam oder als bei der Analyse auszuschließende Kategorie angesehen werden.

Dies kann auch für „weiß nicht mehr", „kann mich nicht erinnern„ oder „verweigert" gelten. Wenn diese Kategorien nicht als Missings definiert werden sollen, erhalten sie einen auf die inhaltlich bereits definierten Kategorien unmittelbar folgenden Code.

- So stellt etwa ein „Weiß nicht" bei Wissensfragen eine inhaltliche Angabe dar und behält deshalb ihren Code als gültigen Wert.

Spezielle Codierungsregeln von Missing Values aufgrund nicht vorgelegter Fragen

Diese Antwortausfälle entstehen durch strukturelle Bedingungen der Erhebung bzw. als Folge von Filterfragen.

- Strukturell bedingte Ausfälle liegen vor, wenn Fragen in einer Welle und / oder in einem Land nicht gestellt wurden. Diese Ausfälle werden gemäß der Standardvorgaben der Erhebung codiert: z. B. durch die Kategorie „not asked in survey" und dem Code „-4" (Beispiel C in Tabelle2) explizit benannt.
- Die Kategorie „trifft nicht zu" („not applicable") umfasst bei Filter-Folgefragen die Befragten, denen die entsprechende Frage nicht vorgelegt wurde, weil sie nicht auf diese Gruppe zutrifft (Filter-non-Response).Ihr wird in der Regel der Code „0" oder ein entsprechender negativer Wert z. B. „-3" zugewiesen (Beispiel A bzw. C in Tabelle 2).

Es können mehrere filternde Kriterien aus Fragen an verschiedenen Stellen des Fragebogens in einer Filterfrage wirksam werden. Generell gilt, dass die filternde Bedingung nur aus dem gültigen Wertebereich der Filterfrage entstammt.

Die Werte, die bereits in der Filterfrage als Missings definiert sind, werden auch in den Folgefragen auf Missing gesetzt.

- Bei der Dokumentation von Filter-Folgefragen sollte für die Kategorie "trifft nicht zu" auch eindeutig angegeben werden, auf welche vorhergehende(n) Codierung(en) der Frage und Kategorie(n) sie sich bezieht. Bei mehrfach hintereinander geschalteten Filtern werden in der Regel alle Filter-Folge-Beziehungen dokumentiert.

Übersicht 5: Codierung einer Filter-Folge-Beziehung

F.17 Waren Sie schon einmal arbeitslos?			F18 (falls Befragter arbeitslos war): Wie lange waren Sie insgesamt arbeitslos?	
1 Ja	n=210	(> F18)	1 Unter einem Jahr	n=150
2 Nein	n=1060	(> F19)	2 Ein Jahr und länger	n= 50
9 KA	n=10	(> F19)	9 KA	n= 10
			0 TNZ (F.17 Code 2oder 9)	n=1070

2.2 Datenkontrolle und Datenbereinigung im Zuge der Datenaufbereitung

Ziel der Datenkontrolle und Datenbereinigung der Originaldaten ist es,

- durch softwaregestützte Verfahren und Techniken systematisch formale und logische Mängel im Datensatz zu identifizieren,
- logische Datenprobleme zu korrigieren, soweit dies sachlich möglich ist,
- formale Mängel der Datendefinition auf Grundlage projektspezifischer Standards anzupassen,
- die Korrekturen zu dokumentieren und die Ergebnisse zu überprüfen.

Die Kontrolle und Bereinigung der Daten, die computergestützt oder durch andere Verfahren erhoben und erfasst wurden, sind eine Voraussetzung, um eine hohe Datenqualität des Analysedatensatzes zu erzielen. Auch eine computergesteuerte Datengewinnung und -verarbeitung garantiert allein noch keine hochwertigen Daten im formal korrekten Sinne. So können Daten zwar formal korrekt erhoben bzw. richtig codiert worden sein, ohne auch logisch konsistent zu sein.

Der Aufwand an Datenkontrollen verringert sich naturgemäß, wenn sie bereits von Erhebungsinstituten vertraglich durchgeführt werden und transparent dokumentiert sind. Dies gilt entsprechend für kontrolliert aufbereitete und dokumentierte Forschungsdaten, die etwa von Datenarchiven oder Forschungsdatenzentren für Sekundäranalysen bereitgestellt werden.

2.2.1 Ursachen für Datenprobleme und Planung der Datenbereinigung

Datenprobleme können aus unterschiedlichen Gründen vorliegen. Sie lassen sich grob in methodische / designbedingte, verhaltensbedingte Gründe sowie formale und technische Ursachen unterteilen. Nur die zuletzt genannte Gruppe ist dabei im engeren Sinne Gegenstand der Datenkontrolle und Datenbereinigung.

Designbedingte Datenprobleme müssen im sozialwissenschaftlichen Methoden- und Analysekontext geprüft und interpretiert werden. So können unpräzise Formulierungen, Positionseffekte in der Abfolge von Fragen, Stichprobenprobleme usw. unter methodischen Gesichtspunkten geprüft oder bereits im Rahmen von Pretests behoben werden. Methodisch durchgeführte Qualitätskontrollen der Feldarbeit dienen wiederum dazu, bewusst falsche Angaben oder Fälschungen ganzer Interviews aufzudecken, was zur kompletten Löschung von Fällen führen kann.

Fehler und Inkonsistenzen, die Gegenstand der rein technisch und formal-logisch motivierten Datenbereinigung sind, können aus unterschiedlichsten Gründen in allen datenbezogenen Arbeitsprozessen während der Datenerhebung, -erfassung und -aufbereitung entstehen. Dazu zählen etwa (vgl. auch Lück, 2011: 74ff)

- Fehler im Erhebungsinstrument bzw. der programmierten Dateneingabemaske, etwa bei der Filterführung oder den Antwortvorgaben;
- Fehler bei der Datenerfassung durch Eingabefehler (falsche Codes oder Zeichen), Spaltenfehler (durch Vertauschen oder Überspringen von Spalten) oder doppelte Erfassung eines Falles bei der manuellen Dateneingabe bzw. Lesefehler beim Scannen ausgefüllter Fragebögen;
- Syntaxfehler in der Steuerdatei oder Fehler in der formalen Datendefinition (Setup), die zu Fehlern beim Einlesen von Rohdaten führen;
- Fehler in der Recordabfolge (Datenzeilen je Befragter) bei einem Datensatz mit mehreren Records je Befragtem;
- Fehler in der Syntax zur Konstruktion von Indizes oder Bildung aggregierter Variablen;
- Fehler durch die Datenbereinigung, etwa im Verlauf der Definition mehrspaltiger Missings oder bei Recodierungen von Variablenwerten.

Missverständnisse zwischen Interviewer und Befragten oder Irrtümer bei der Beantwortung von Fragen können je nach Untersuchungssituation eventuell noch durch Rücksprache mit Interviewern bzw. Befragten und durch die Kontrolle der Erhebungsunterlagen behoben werden.

Deshalb sollte ein Forschungsprojekt relativ frühzeitig konkrete Leitlinien für den Umgang mit diesen typischen Datenproblemen formulieren und operative Regeln für die Bereinigung von Datenproblemen in der Datenaufbereitung festlegen (vgl. 2.2.2).

Mit der Planung dieser konkreten Arbeitsschritte ist weiterhin zu überlegen, wie die Modifikationen von Variablenwerten des Originaldatensatzes während der Datenaufbereitung dokumentiert werden sollen. In diesem Kontext ist insbesondere zu entscheiden, ob Routinekontrollen und spezifische Modifikation einzelner Variablen menügestützt oder mit Hilfe von Steuerbefehlen des Datenanalyseprogrammes durchgeführt werden sollen.

Ein wesentliches Entscheidungskriterium, das für eine skriptbasierte Datenprüfung (im Unterschied zu einem menübasierten Vorgehen) spricht, ist die dadurch erzielte Transparenz, Nachvollziehbarkeit und Reproduzierbarkeit der Schritte und Ergebnisse der Datenbereinigung.

Der Vorteil, allgemeine und spezifische Prozeduren jeweils in besonderen Syntaxdateien zu definieren und (auf Vorrat) zu speichern, liegt auch in der Standardisierung (und einfachen Modifikation) wiederholt erforderlicher Prozeduren in Erhebungsprojekten. Um die Datenbereinigung, deren Zeitaufwand oftmals unterschätzt wird, zu beschleunigen, ist es weiterhin empfehlenswert, frühzeitig einen Satz wichtiger Prüfroutinen des eingesetzten Statistikprogramms zu erstellen und sie zusammen mit dem Pretest oder parallel zur eigentlichen Datenerhebung zu testen.

Indem die Bearbeitungsschritte in versionierten Syntaxdateien gesichert und auf diese Weise dokumentiert werden, können die durchgeführten Arbeitsschritte während der Datenaufbereitung transparent nachvollzogen werden. Damit erleichtern sie auch das Erkennen von Fehlern, die etwa bei Recodierungen oder der Bildung abgeleiteter Variablen entstehen können.

Ein zusätzliches Hilfsmittel stellen in diesem Zusammenhang auch Logfiles dar, die allerdings nicht als Standardfunktionen in allen Statistikprogrammen zur Verfügung stehen. Je nach Programmkontext dokumentieren sie unmittelbare Änderungen an den Daten, z. B. wenn in Stata Daten manuell im Dateneditor verändert werden. SPSS wiederum bietet die Möglichkeit, mit Hilfe des Viewers Prozeduren, die menübasiert gestartet wurden, zu loggen und sie zusammen mit den Ergebnissen in verschiedene externe Formate zu exportieren.

Zusammenfassend erscheint es im Interesse von Transparenz, Nachvollziehbarkeit und Reproduzierbarkeit der geplanten Abläufe und Arbeitsschritte in der Datenaufbereitung und Datenanalyse sinnvoll und notwendig, bereits vor der Datenerhebung die

- zu verwendende Software bzw., Statistikprogramm (-module) unter Berücksichtigung des erforderlichen Funktionsumfangs zur Erhebung, Bereinigung und Analyse der Daten festzulegen;
- Anforderungen an Abläufe und Routinen zur Konvertierung oder zum Austausch von Daten bzw. Dateiformaten zwischen Anwendungen (Import, Export, Dateiformat) auf Funktionalität und den erforderlichen Aufwand hin zu prüfen.

In diesem Zusammenhang ist dann auch die Organisation der anfallenden Dateien projektspezifisch zu planen. Dies betrifft alle

- Arten von Datenfiles (Originaldaten, Arbeitsdateien), Steuerungsdateien (Syntaxfiles) und Ergebnisdateien, die auf der Grundlage von Projektkonventionen logisch organisiert werden müssen (Dateinamen, Verzeichnisse);
- Veränderungen, die durch (Datei-)Versionen gekennzeichnet werden sollten;
- technischen und administrativen Verfahren zur Datensicherung und zum Datenschutz.

Weitere Hinweise zu den verschiedenen Aspekten der Datenorganisation sind in Kapitel 3 beschrieben.

2.2.2 Einzelschritte der Datenkontrolle und Datenbereinigung

Grundsätzlich sollte der zu erstellende Analysedatensatz alle verfügbaren empirischen Informationen aus der Erhebung enthalten. Der Detaillierungsgrad der Daten sollte dem der erhobenen Originaldaten entsprechen. Das im Weiteren beschriebene Vorgehen folgt den Empfehlungen zur Datenkontrolle und Datenaufbereitung im GESIS Datenarchiv (Brislinger et al. 2009).

Hinsichtlich der Datenanalysen und der Beschreibung der Ergebnisse ist erstens zu überlegen, wie die während der Datenaufbereitung und Datenbereinigung vorgenommenen Korrekturen am Originaldatensatz dokumentiert werden. Zweitens ist zu klären, wie die eventuell im Datensatz (noch) vorhandenen Probleme erörtert und dokumentiert werden sollen.

Für den Umgang mit Datenproblemen sollte das Projekt klare Regeln definieren und dokumentieren. Dabei ist grundsätzlich zu entscheiden, ob überhaupt in die Daten eingegriffen werden soll. Dann sollte festgelegt werden,

- ab welcher Größenordnung (Anzahl der Fälle) Probleme in den Daten dokumentiert, aber nicht recodiert werden;
- ab welcher Größenordnung (Anzahl der Fälle) Probleme durch Recodierung behoben werden (schließt Dokumentation ein), z. B. durch
  - Recodierung (mit Hilfe anderer Variablen) in inhaltliche Werte,
  - Recodierung in „keine Angabe" oder „trifft nicht zu";
- ob nicht-plausible aber logisch mögliche Werte recodiert und dokumentiert oder nur dokumentiert werden (Letzteres ist eher zu empfehlen);
- ob Widersprüche zwischen verschiedenen Antworten eines Befragten recodiert und dokumentiert oder nur dokumentiert werden (Letzteres ist eher zu empfehlen);
- ob Missing Values (z. B. „keine Angabe") auf der Basis der Informationen anderer Variablen rekonstruiert und durch inhaltliche Codes ersetzt werden sollen.

Der Prozess der Datenaufbereitung und der Datenkontrolle ist angesichts der unterschiedlichsten Detailfragen kein starrer, linearer Prozess, sondern ein iteratives Vorgehen und Ineinandergreifen von verschiedenen Einzelschritten, die sich teilweise gegenseitig bedingen. Es können jedoch die folgenden Vorgehensweisen und Regeln der formalen und logischen Datenkontrollen beschrieben werden, wie sie routinemäßig von sozialwissenschaftlichen Forschungsprojekten, Erhebungsinstituten oder Datenarchiven bei der Datenaufbereitung eingesetzt werden.

Formale Prüfungen zum Auffinden von Fehlern und anschließende Korrekturen

Vor jeder Kontrolle und Recodierung der Daten sollte darauf geachtet werden, dass in der Auszählung die volle Fallzahl erreicht wird. Insbesondere sollte allen System Missings (leere Missing Values) ein konkreter Wert (User Missing Values) zugewiesen werden, auch um die volle Fallzahl in einer Grundauszählung zu erreichen und kontrollieren zu können. Weiterhin werden folgende Verfahren zum Auffinden und Korrigieren von formalen Datenfehlern angewendet:

- Prüfen der Befragten-ID auf Eindeutigkeit:
  Fälle mit gleicher ID werden daraufhin überprüft, ob sie doppelt erfasst oder die ID mehrfach vergeben wurde. Im Ergebnis erhält man die Anzahl gültiger Fälle, die für spätere Kontrollen eine wichtige Bezugsgröße darstellen.

Entsprechend gilt die Eindeutigkeitsprüfung für ein Land oder eine Welle, wobei die effektive Stichprobengröße mit der Anzahl der Anzahl der Fälle je Land / Welle übereinstimmen muss.

- Vergleich der Variablen auf formale Übereinstimmung mit den Fragen und Antwortkategorien im Fragebogen:
  Der Codeplan, die Variablenliste und der Fragebogen stellen die dazu nötigen Informationen bereit. Die einwandfreie und konsistente Datendefinition erleichtert die weiteren logischen Kontrollen der Daten.
- Besondere Aufmerksamkeit sollte dabei auf die Überprüfung von Variablenwerten auf „Wilde Codes" gelegt werden:
  Dies sind Werte, die außerhalb des definierten bzw. zugelassenen Wertebereiches liegen. Eine systematische Prüfung ist am einfachsten, wenn die Häufigkeiten aller Variablen ausgezählt werden und alle Value Labels in den Outputs erscheinen. Dadurch lassen sich Abweichungen einfach erkennen.
- Überprüfung der definierten Spaltenformate der Variablen:
  z. B. Dezimalpunkt und Dezimalstellen bei Gewichtungsvariablen, Formate bei mehrspaltigen Missing Kategorien. Fehlerfreie Formate der Variablen sind eine wesentliche Grundlage für nachfolgende Datenkontrollen und natürlich für die Analysen im entsprechenden Softwarepaket.
- ‚Sichtkontrolle' der Daten:
  Diese Form der Kontrolle richtet das Augenmerk auf leere Felder oder auffälliger ‚Verschiebungen' von Werten in den Spalten der Datenmatrix. Sie dient im Zusammenhang mit der Häufigkeitsauszählung aller Variablen der systematischen Vorprüfung aller Variablenwerte und der Nachkontrolle aller Datendefinitionen.

Konsistenzkontrollen

Bei dieser Art von Kontrollen ist zu prüfen, ob sich Werte in verschiedenen Variablen aus formalen und / oder inhaltlichen Gründen explizit bzw. implizit widersprechen. Dabei lassen sich folgende Verfahren unterscheiden:

Überprüfen der formalen Konsistenz von Filterführungen hinsichtlich

- aller Variablen, die in die Filterführung im Fragebogen einbezogen sind und der
- Einhaltung der vorgegebenen Filterbedingungen.

  Die Summe aller nicht zu befragenden Fälle muss mit „Trifft nicht zu" codiert sein. Sie sind abzugrenzen von den Fällen, in denen ein im Sinne der Filterführung zulässiger Befragter „keine Angabe" zur Folgefrage macht.

Prüfen der Antworten bei Variablen mit Mehrfachnennungen auf Konsistenz hinsichtlich

- der Einhaltung der Antwortbegrenzungen (z. B. max. 3 Antworten) und Dokumentation von Abweichungen und
- Prüfung der Kategorie/Variable für „Keine Antwort auf alle Items der Frage", falls vorhanden.

Prüfen der Konsistenz auf inhaltliche Fehler durch

- Identifizierung logisch ausgeschlossener Beziehungen zwischen Variablen auf Fallebene (‚schwangerer Mann") unter Anwendung eigener Filterbedingungen oder durch Kreuztabellierung logisch zusammenhängender Variablen.

Die durchgeführten Änderungen der Daten sollten in einem neuen (versionierten) Datenfile gespeichert werden. Diese Daten sind dann erneut auf mögliche Fehler; die durch die Modifikation entstanden sein können, zu prüfen. Die geprüften (endgültigen) Datenfiles sowie die Syntaxdatei(en) zu ihrer Erzeugung werden abschließend unter entsprechend aussagekräftigen Dateinamen abgespeichert.

Plausibilitätskontrollen

Ziel ist es, die Daten daraufhin zu prüfen, ob ein Wert stimmig, realistisch, nachvollziehbar oder glaubwürdig ist bzw. sein kann. Dabei ist die Schwierigkeit zu berücksichtigen, dass Werte unplausibel oder unwahrscheinlich erscheinen mögen, aber logisch möglich sein können (z. B. eine Familie mit 14 Kindern).

- Prüfen der demographischen Variablen auf Aggregatebene:
  z. B. durch Vergleich der Verteilung im Datensatz mit einer Bevölkerungsstatistik hinsichtlich Geschlecht, Alter, Familienstand, Einkommen, Kinder im Haushalt, Region.

  Der Identifizierung und Überprüfung von Ausreißern (Extremwerte) bei Variablen ohne definierten Wertebereich gilt dabei besondere Aufmerksamkeit. Dies betrifft z. B. extrem hohe (oder niedrige) Angaben zum Alter, zum Einkommen, zur Anzahl von Kindern usw. aber auch Konstellationen, wie eine 14 jährige Abiturientin, einem verheirateten Zehnjährigen usw.

- Prüfen der Variablen Gewichte (erwarteter Mittelwert =1).

Zur Beantwortung der Frage, ob ein Wert realistisch ist, liefern sowohl Kontextinformationen aus anderen Variablen und Erhebungsunterlagen, aber auch Alltags- und Expertenwissen sowie statistische Informationen und weitere inhaltlich relevante Quellen oftmals zusätzlich benötige Informationen, um die Frage aufzuklären. Beim Prüfen von Variablen auf Aggregatebene können neben dem Vergleich mit unabhängigen Referenzdaten auch vergleichbare Ergebnisse aus anderen Wellen (Länder, Zeitpunkte) einbezogen werden.

Je nach Situation bietet es sich an, unplausible aber logisch mögliche Werte oder etwa eine hohe Anzahl von Missing Values bei heiklen Fragen (zum Mogeln bei Steuererklärungen, nach Diebstählen oder Sexualkontakten etc.) zunächst nur zu dokumentieren, um diese Werte durch weitere statische Datenanalysen z. B. hinsichtlich eines nicht-zufällig bedingten Antwortverhaltens genauer zu untersuchen.

3 Organisation und Sicherung der Daten und Dokumente

Je nach Forschungstyp, Umfang und Art der erhobenen Daten sowie der Art ihrer technischen Organisation, Bearbeitung und Speicherung (Datei; Datenbank) entstehen unterschiedliche Formen und Sets von Einzeldateien in einem Forschungsprojekt. Dazu zählen insbesondere

- Informationen zur Fragebogenentwicklung, zu Pretests und Untersuchungsdesign
- Originalfragebogen und Stimuli, Codeplan, Feldbericht
- Rohdatenfile mit den Originaldaten der Erhebung,
- Arbeitsdateien sowie Syntax- / Steuerdateien zur Datenaufbereitung,
- Analysedatenfiles und Syntaxdateien zu statistischen Datenanalysen,
- Datenoutputs (Häufigkeiten, Kreuztabellen, Visualisierung von Datenanalysen).

Diese in Dateien erfassten Informationen werden

- für verschiedene Zwecke (Datenbearbeitung, Projektdokumente, Dokumentationen),
- in unterschiedlichsten Formen (Originaldatei, Masterdatei, Arbeitsdateien),
- zu unterschiedlichsten Zeiten (temporäre Arbeitsdatei; Entwurf, Zwischenversion, Endversion) erstellt, bearbeitet und eventuell an unterschiedlichen Orten abgelegt.

Für alle Dateien, die im Projektbetrieb erstellt werden, sind qualitätssichernde Maßnahmen zu deren Organisation, Sicherung und Bereitstellung zu planen und als eigenständiger Teil im Datenmanagementplan zu beschreiben. Naturgemäß erfordert der Umgang mit den erhobenen Forschungsdaten im Zuge der Datenorganisation und -verarbeitung im Projektbetrieb dabei besondere Beachtung.

Projektspezifischen Regeln und Verfahren zur logischen Organisation und technischen Sicherung der Daten und Dateien sowie die entsprechenden personellen Verantwortlichkeiten und technischen Maßnahmen sollten im Interesse der Qualitätssicherung und Transparenz der Vorgaben in geeigneter Form z. B. als Bestandteil eines Projektleitfadens zum Datenmanagement und / oder im Datenmanagementplan schriftlich festgehalten werden.

Bei der Organisation und Sicherung der Daten kann zwischen logischen (Abschnitt 3.2) und technischen Aspekten (Abschnitt 3.3) unterschieden werden. Zunächst werden jedoch die Verfahren zur Gewährleistung der Datensicherheit von Maßnahmen im Zusammenhang mit dem Datenschutz im folgenden Abschnitt abgegrenzt.

3.1 Datensicherheit und Datenschutz

Die Datensicherheit betrifft alle technischen und organisatorischen Maßnahmen zum Schutz der physischen Daten vor ungeplanter Veränderung sowie vor Verlust und Zerstörung durch menschliche und technische Fehler sowie durch Missbrauch oder höhere Gewalt.

Der Datenschutz behandelt demgegenüber die rechtlichen Aspekte zum Schutz der Persönlichkeit von Befragten u. a. im Zusammenhang mit der Verarbeitung personenbezogener Daten. Sie werden im Bundesdatenschutzgesetz (BDSG), in Länderdatenschutzgesetzen sowie weiteren bereichsspezifischen Regelungen reguliert. Die Anwendung eines entsprechenden Gesetzes ist davon abhängig, ob die Datenverarbeitung in einem öffentlichen oder privaten Bereich stattfindet.

Datenschutzrechtliche Aspekte und Regelungen im Zusammenhang mit sozialwissenschaftlichen Forschungsvorhaben wurden bereits in Abschnitt 1.3 unter dem Aspekt der Planung und Durchführung einer Datenerhebung aufgegriffen. Im Folgenden wird auf technisch-organische Maßnahmen hingewiesen, die bei der Verarbeitung personenbezogener Daten zu berücksichtigen sind.

Nach § 9 BDSG müssen öffentliche und nicht-öffentliche Stellen, die selbst oder im Auftrag personenbezogene Daten verarbeiten, technische und organisatorische Schutzmaßnahmen treffen, um die Umsetzung der Anforderungen des Bundesdatenschutzgesetzes zu gewährleisten. Der Aufwand für diese Maßnahmen muss in einem angemessenen Verhältnis zum angestrebten Schutzzweck stehen. Die in der Anlage zu § 9 Satz 1 BDSG beschriebenen Maßnahmen sollen - je nach Art der zu schützenden personenbezogenen Daten oder Datenkategorien - geeignet sein,

- die Kontrolle des Zutritts, Zugangs und Zugriffs auf Datenverarbeitungssysteme
- sowie die Kontrolle der Weitergabe, Eingabe, Auftragsverarbeitung, Verfügbarkeit und die getrennte Verarbeitung von Daten, die für unterschiedliche Zwecke erhoben wurden, zu gewährleisten.
- Gemäß § 4f BDSG müssen öffentliche und nicht-öffentliche Stellen, z. B. auch Forschungseinrichtungen und Erhebungsinstitute einen Datenschutzbeauftragten schriftlich bestellen.

In Abhängigkeit von der jeweiligen institutionellen Einbindung des Forschungsvorhabens (Universität, Forschungseinrichtung etc.) ist im Detail zu klären, welche administrativen Regelungen und technischen Infrastrukturen hinsichtlich der Datensicherheit und zum Datenschutz vorhanden und zu berücksichtigen sind. Die Rechte des Datenzugangs für Projektmitarbeiter unter besonderer Berücksichtigung des Zugangs zu und der Bearbeitung von personenbezogenen und sonstigen sensiblen Daten und Informationen sind auch administrativ festzulegen. Dabei ist zu klären, wer im Projekt auf welche Daten zugreifen kann und wer welche Dateien verändern und / oder löschen darf. Angaben über entsprechende Verfahren und Maßnahmen sollten in den Datenmanagementplan und / oder den Projektleitfäden zum Datenmanagement einfließen.

3.2 Logische Aspekte und Konventionen zur Dateiorganisation

In diesem Zusammenhang ist festzulegen, wie die unterschiedlichsten Dateien und Dateiformen formal in einem Projekt gehandhabt werden sollen. Dazu zählen insbesondere die Konventionen zur Benennung von Verzeichnisstrukturen und Dateien sowie die Versionierung der relevanten Daten- und Dokumentationsdateien. Weiterhin sind die Zugriffsrechte der Projektmitarbeiter auf die Dateien administrativ zu regeln.

Konventionen zur Organisation von Dateien und Verzeichnissen

Eindeutige Namenskonventionen helfen, Dateien einfach zu erkennen, zu finden und physikalisch leicht zugreifbar zu halten. Entsprechend sollten im Voraus Regeln für die (hierarchische) Organisation von Dateien in Verzeichnissen und für die strukturierte Bezeichnung von Dateien und Verzeichnissen entsprechend des Projektbedarfes zur jeweiligen Projektphase festgelegt werden.

Zur Gewährleistung von Kompatibilität zwischen MacOS, Unix und Windows sollten folgende Beschränkungen bei der Nutzung von Zeichen und Sonderzeichen in Dateinamen und Verzeichnisnamen berücksichtigt werden:

- Zeichen: a .. z, A .. Z, 0 .. 9 (max. 31 Zeichen bei Dateinamen).

- Als Sonderzeichen sollten nur Unterstriche (_) oder Bindestriche (-) verwendet werden. Unterstriche ersetzten Blanks und Bindestriche werden innerhalb von Namen benutzt. Es sollten also keine Leerzeichen und keine deutschen Umlaute in Dateinamen und Verzeichnisnamen benutzt werden.

Je nach Umfang und Komplexität der zu bearbeitenden Dateien kann es hilfreich sein, getrennte Ordnerstrukturen für Rohdaten, Analysedaten und Datenauswertungen sowie weitere Projektmaterialien zu definieren.

Arten von (Daten-)Files und Namenskonventionen für Dateinamen

Nach der Datenerhebung werden die gewonnenen Originaldaten während der Datenaufbereitung formalen und logischen Kontrollen unterzogen und die Daten bereinigt. Mit der Entwicklung eines Analysefiles unterliegen Daten typischerweise weiteren Anpassungen, etwa durch die inhaltliche Harmonisierung von Variablen oder die Konstruktion von zusätzlichen Variablen. Um diese schrittweise erzeugten Datenmodifikationen auch auf Dateiebene transparent und reproduzierbar zu verwalten, sollten die zusammengehörenden Dateien entsprechend versioniert werden.

Grob lassen sich folgende Arten von Datenfiles voneinander abgrenzen:

1. Der Originaldatenfile: Die Originaldaten der Datenerhebung werden schreibgeschützt gesichert und nicht mehr verändert. Datenanpassungen werden an Kopien der Masterdatei vorgenommen und mit Versionsangaben (schreibgeschützt) als Arbeitsdateien gespeichert.
2. Der bereinigte Datenfile: Nach Abschluss aller formalen Datenkontrollen und entsprechenden Korrekturen liegt ein bereinigter Datensatz vor. Auch diese Datei wird schreibgeschützt gesichert und nicht mehr verändert. Weitere Schritte der Datenaufbereitung (z. B. Bildung weiterer Variablen) werden an Kopien dieser Datei durchgeführt und als Versionen gespeichert.
3. Der Analysefile: Die erste Version des analysefähigen Datenfiles wird schreibgeschützt gespeichert. Analytische Auswertungen und methodengeleitete Modifikationen der Daten können an (Arbeits-) Kopien vorgenommen und auf Basis weiterer Konventionen gespeichert werden.

Die unterschiedlichen Arten von Datenfiles und die Veränderungen an den Daten können etwa durch Angaben zu Typ und Version im Dateinamen kenntlich gemacht werden. Parallel müssen die Änderungen in geeigneter Weise dokumentiert werden (Textdokument; Syntaxfile).

Neben den datenbezogenen Dateien sind auch weitere Dokumente wie etwa Codeplan, Methodenberichte; Auswertungen, usw. mit aussagekräftigen Namen (und Typbezeichnung) zu benennen und zu versionieren. Dabei ist es wichtig, zwischen verschiedenen Dateiformen zu unterscheiden, die in den unterschiedlichen Phasen der Datenaufbereitung und im Verlauf der Datenanalysen entstehen können:

- Ausgangsdateien, z. B. Originaldaten; Codeplan, etc.,
- Arbeitsdateien (temporäre Datei; Entwurf, Zwischenversion) sowie
- Ergebnisdateien (Bericht, Report / Syntax und bereinigter Datenfile bzw. Analysefile).

Versionierung von Datensätzen

Generell sollten Änderungen am Datensatz und den Metadaten des Statistikprogramms (Syntaxdatei mit Datendefinition, Labels etc.) zusammen nach einheitlichen Konventionen gespeichert werden. Entsprechende Verfahren sind vom Projekt festzulegen. Sie sollten auch ein Versionierungskonzept mit entsprechenden Versionsnummern zumindest für die zu bearbeitenden Datenfiles einschließen. Dazu

kann ein Nummernkonzept eingesetzt werden, wie es von Datenarchiven auf Grundlage des DDI-Standards benutzt wird. Die drei Bestandteile des Konzeptes sind:

- „Major.Minor.Revision".
- Die Major-Nummern beginnen mit 1. Minor- und Revision-Nummern beginnen immer mit 0.

Die erste Version eines Datenfiles heißt demnach „1.0.0". Die Nummern werden bei Datenänderungen gemäß der Relevanz für die Aussagekraft der Daten anhand folgender Kriterien verändert. Die Anzahl von Ebenen der Nummerierung und ihre Parameter können an den jeweiligen Bedarf des Projektkontextes angepasst werden.

Übersicht 6: Drei-Ebenen Versionierung und Kriterien zur Anpassung der Versionsnummer

- 1. Position / Major-Nummer: Sie wird verändert, wenn
  - eine oder mehrerer Fälle in (aus) einem Datensatz eingefügt (gelöscht) werden;
  - eine oder mehrerer Variablen in (aus) einem Datensatz eingefügt (gelöscht) werden;
  - eine oder mehrere neue Wellen in einen integrierten Datensatz eingefügt werden;
  - ein o. mehrere Sample in den integrierten o. kumulierten Datensatz eingefügt werden.
- 2. Position / Minor-Nummer: Sie wird verändert, wenn Änderung einer Variablen, d. h. bedeutungsrelevante Korrekturen oder Ergänzungen im Datensatz vorgenommen werden (Labels, Recodierungen, Datenformate, etc.).
- 3. Position / Revision-Nummer: Sie wird verändert, wenn etwa einfache Überarbeitungen von Labels ohne Bedeutungsrelevanz vorgenommen werden.

Die jeweilige Versionsnummer sollte zur einfachen Erkennung der Dateien im Filesystem in den Dateinamen eingefügt werden, z. B. XYZ_v1-1-0.sps.

Datensatzkorrekturen können zusätzlich in einer Versionshistorie außerhalb des Datenfiles als Übersicht zu noch bestehenden Problemen oder offenen Arbeitsschritten dokumentiert werden. Je nach Komplexität des Datensatzes und /oder der Datenbearbeitung in verteilten Forschungsteams ist zu überlegen, ob die Versionsnummer durch eine Variable im Datensatz dokumentiert werden soll. In einem solchen Fall sind folgende Parameter festzulegen:

Übersicht 7: Definition einer Versionsvariablen im Datensatz

- Variablenname: Er sollte sich grundsätzlich an den Projektkonventionen zur Namensvergabe orientieren (V-Nummer, Mnemo, ...). Default-Name der Mnemo-Variante wäre VERSION.
- Das Variablenlabel verweist auf die versionierende Person, Gruppe oder Institution.
- Variablentyp/-format: STRING, Spaltenbreite '25'.Das STRING-Format ist wichtig, um auch bei zwei- (oder mehr-) stelligen Werten auf den einzelnen Positionen durch Trennzeichen eine klare Unterscheidung der Ebenen sicherzustellen.
- Value: Der Datenwert selbst setzt sich aus der Versionsnummer, einem folgenden Leerzeichen (blank) sowie dem Versions-Datum in Klammern im ISO-Format zusammen: „Major.Minor.Revision (YYYY-MM-DD)" – Beispiel: „2.1.10 (2010-03-25)".
- Position im Datensatz: Standardmäßig sollte die Versionsvariable im Anfangsbereich des Datensatzes zu weiteren administrativen hinzugefügt werden.

3.3 Technische Aspekte und Konventionen der Dateiorganisation

Hierzu zählen die Verfahren und Maßnahmen, durch die Datenfiles und alle anderen Projektdateien technisch gespeichert und für den Zugriff verfügbar gehalten werden. Dabei ist administrativ zu bedenken wie lange (temporär, Projektlaufzeit, Langzeitverfügbarkeit) die Dateien in welchem Format und an welchem Ort gespeichert werden sollen und wie die unterschiedlichen Zugriffsrechte technisch realisiert werden.

Technische Plattform für den Datenaustausch und die Datensicherung

Im Zusammenhang mit der Anzahl von Mitarbeitern, Partnern und Standorten eines Projektes ist zu überlegen, wie die technische Plattform aufgebaut sein muss, um Daten zwischen den Mitgliedern eines Forschungsprojektes auszutauschen und sicher zu speichern. Auch wenn dies heutzutage kein generelles Kostenproblem darstellt, müssen auch die erwarteten Datenmengen bei der Planung einer solchen Plattform berücksichtigt werden.

Bei kleinen Projekten kann etwa ein gemeinsam genutzter Fileserver ausreichend sein. Mit wachsender Größe und unterschiedlichen Standorten kann die Zusammenarbeit etwa über webbasierte Projektportale (Repository; File Sharing System) organisiert werden.

Unabhängig von der Projektgröße sind angemessene technische Maßnahmen zur datenschutzkonformen Sicherung von personenbezogenen Daten einzuplanen, z. B. in Form von Zugangskennwörtern sowie Schreib- und Leserechten für die benutzten Verarbeitungssysteme. Werden institutionelle Netzwerkspeicher genutzt, sind die entsprechenden Richtlinien in die Planung des Datenmanagements einzubeziehen.

Datei- und Datenformate

Die Frage, in welchen technischen Dateiformaten Daten, Projektdokumente oder Metadaten erfasst, unterhalten, gespeichert und verfügbar gemacht werden sollen, sollte unter Berücksichtigung etablierter Standards in den Arbeitsroutinen der Sozialwissenschaften betrachtet werden.

Aus pragmatischen Gründen wird während der Projektlaufzeit die Verwendung eingeführter und verbreiterter Software (z. B. SPSS, STATA, SAS oder R für die Datenbearbeitung) und entsprechender (proprietärer) Dateiformate üblich sein. Gleichzeitig gewinnt der Einsatz offen dokumentierter Dateiformate an Bedeutung, z. B. XML basierte Formate in der Datendokumentation, die auch im Interesse einer nachhaltigen Sicherung der Daten vorteilhaft eingesetzt werden können.

Um einen relativ reibungslosen Austausch zwischen verschiedenen Systemen zu ermöglichen, sollte der Einsatz eines möglichst homogenen Sets an Programmen und Dateiformaten geplant werden. Dies vermeidet Zeit- und Informationsverluste durch Konvertierungen der Dateien und verringert Anforderungen an die Datenorganisation. Standardroutinen zur Konversion und zum Austausch von Daten und Metadaten zwischen unterschiedlichen Bearbeitungsprogrammen sind gegebenenfalls zusätzlich zu berücksichtigen.

In Hinsicht auf eine spätere Archivierung dürften in der Regel keine Probleme bei der Handhabung von Dateiformaten aktuell verbreiteter Software durch datenhaltende Einrichtungen, Forschungsdatenzentren oder Datenarchive bestehen.

Dateispeicherung und Datensicherung durch Backups

Bei der Planung und Umsetzung entsprechender technischer Maßnahmen sollten die verschiedensten Gründe von Daten- und Dateiverlusten berücksichtigt werden, z. B. Hardwareversagen, Softwarefehler, Viren, Stromausfälle oder menschliche Fehler.

Vor dem Erstellen der ersten wichtigen Daten und Dateien sollten die wesentlichen technischen und administrativen Aspekte zur Datensicherung und Datensicherheit im Projekt geregelt sein. Gleichzeitig ist zu empfehlen, die technischen Maßnahmen zur Datensicherung verbindlich zu dokumentieren. Auch automatisierte Verfahren sollten verschriftlicht werden, um die Transparenz der Abläufe zu gewährleisten und die zu sichernden Objekte eindeutig zu kennzeichnen.

Die folgenden Maßnahmen zur Sicherung aller projektrelevanten Dateien (auf Einzelplatz-PCs bis hin zu zentralen Projektressourcen) können gegebenenfalls bei der Planung des projektspezifischen Vorgehens berücksichtigt bzw. eingesetzt werden:

Übersicht 8: Maßnahmen zur Sicherung projektrelevanter Dateien

- Einsatz aktueller Antivirensoftware und sonstiger Software zum Schutz vor Schadcode.
- Festlegung der zu sichernden Daten und Dateien (Originaldateien, Masterdateien, Arbeitsdateien etc.) bzw. des gesamten Systems (z. B. Festplattenimage o. ä.).
- Bestimmung der Anzahl an Kopien und deren Ablageort, z. B. als Doppelstrategie mit Sicherung auf Festplatte und Netzlaufwerk oder DVD.
- Festlegung der physikalischen Speichermedien, -formate und -orte (On-line / Off-line), auf denen die zu sichernden Dateien abgelegt werden.
- Häufigkeit und Art der Datensicherung, z. B. nach jeder Datenänderung und in festen Zeitabständen als vollständige, inkrementelle oder differentielle Sicherung.
- Mindestens ein Satz der Daten sollte nicht verschlüsselt gesichert werden, um die technische Wiederherstellung bzw. Rekonstruktion der Dateien nicht zusätzlich zu erschweren.
- Regelung der Aufbewahrungsfristen und Regeln zum Löschen von Dateien (falls notwendig).
- Festlegung der technischen Zugriffsrechte (Passwörter, Firewall, Lese- und Schreibberechtigungen) und Schutz vor Überschreibung eines Backup Satzes (Read-only).
- Je nach Projektdauer sollten regelmäßig die Datenintegrität und die Vollständigkeit aller gesicherten Dateien geprüft werden.

4 Metadaten und Standards zur Studien- u. Datendokumentation

Standardisierte Metadaten dienen der nachhaltigen Dokumentation, Sicherung und Nachvollziehbarkeit von Forschungsdaten. Gleichzeitig ermöglichen sie ein dauerhaftes Auffinden und Verständnis datenbasierter Forschungsergebnisse und die Zitation der zugrundeliegenden Forschungsdaten. Verstanden als „Daten über Daten" sind Metadaten Informationen,

> „die in strukturierter Form analoge oder digitale Forschungsdaten (Objekte) ... beschreiben, erklären, verorten oder definieren ... (NISO, 2004). Im Folgenden werden Forschungsdaten vereinfacht als das Objekt verstanden, das von einer übergeordneten Ebene aus mit Metadaten beschrieben wird." (Jensen, Katsanidou, Zenk-Möltgen 2011: 83)

Die für sozialwissenschaftliche Forschungsvorhaben interessanten Metadatensysteme und Standards werden in den folgenden Abschnitten vorgestellt (Quellen vgl. Anhang A.3). Zunächst wird auf einige Beispiele wichtiger Normen und Klassifikationssysteme eingegangen, die fachliche Standards darstellen, die bei der sozialwissenschaftlichen Erhebung und Codierung spezifischer Merkmale genutzt werden, z. B. als Standarddemographie oder bei der Codierung von Berufen (Abschnitt 4.1).

Anschließend wird das DDI-Metadatensystem vorgestellt, das seit längerem als technischer Standard zur Dokumentation sozialwissenschaftlicher Forschungsdaten und Studieninformationen etabliert ist (Abschnitt 4.2; Quellen Anhang A.4).

Die Möglichkeit zur dauerhaften Zitation von Sozial- und Wirtschaftsdaten durch Persistent Identifier und das damit verbundene Metadatenschema wird in Zusammenhang mit der Datenregistrierungsagentur da|ra im letzten Abschnitt 4.3 thematisiert.

4.1 Sozialwissenschaftlich relevante Standards und Klassifikationen

Sozialwissenschaftliche Standards und Skalenhandbücher sind eine wichtige Informationsquelle bei der Planung und der Entwicklung sozialwissenschaftlicher Erhebungsinstrumente. Um die Vergleichbarkeit von Daten zu erleichtern, werden außerdem spezifische Merkmale wie Berufe oder Ausbildung anhand von nationalen oder internationalen Klassifikationssystemen codiert. Darüber hinaus werden bei der standardisierten Dokumentation von Länder- oder Sprachangaben entsprechende ISO-Normen genutzt.

- **ZIS** (2010) ist eine Zusammenstellung sozialwissenschaftlicher Items und Skalen. Das elektronische Handbuch ZIS dokumentiert Instrumente zur Erhebung von Einstellungen und Verhaltensweisen aus häufig untersuchten sozialwissenschaftlichen Themengebieten, gemeinsam mit zentralen theoretischen und methodischen Ansätzen.
- **EHES** (2010) ist das Elektronische Handbuch zu Erhebungsinstrumenten im Suchtbereich. Das Handbuch unterstützt empirische Untersuchungen über Einstellungen und Verhalten im Suchtbereich. Dazu werden Befragungsinstrumente aus dem Suchtbereich gemeinsam mit einschlägigen theoretischen und methodischen Informationen und Daten zur Beurteilung ihrer Güte dokumentiert.

Die Beiträge zu den Instrumenten in ZIS und EHES werden von Fachautoren verfasst und als wissenschaftliches Dienstleistungsangebot von GESIS Leibniz-Institut für Sozialwissenschaften veröffentlicht.

- Die **Demographischen Standards** (2010) verstehen sich als Empfehlungen zur Vereinheitlichung von sozialstrukturellen Erhebungsmerkmalen in deutschen Haushalts- und Personenbefragungen. Neben den 20 Kernvariablen (von Alter bis zu Einkommensabfragen) wird u. a.

auf Typisierungen, Indizes sowie Klassifikation von Berufen eingegangen. Weiterhin wird die Handhabung und praktische Anwendung demographischer Standards anhand konkreter Fragetexte mit Antwortkategorien und Auswahllisten (Download siehe Anhang A.3) in unterschiedlichen Erhebungsformen thematisiert. Ziel dieser Standards ist es, die Vergleichbarkeit von Forschungsdaten aus Umfragen zu erhöhen.

Weiterhin werden nationale und internationale Klassifikationssysteme zu Berufen und Ausbildungssystemen in sozialwissenschaftlichen Untersuchungen zur standardisierten Erhebung, Codierung und Dokumentation soziodemographischer Merkmale angewendet.

- **ISCED** ist eine Klassifikation von Schulsystemen und Schultypen, die von der UNESCO entwickelt wird (1997; Rev. 2011). Das UNESCO Institute for Statistics stellt aktualisierte Zuordnungen auf Länderebene z. B. für Deutschland zur Verfügung (ISCED Mappings 2011).
- **ISCO**, die Internationale Standardklassifikation von Berufsgruppen, wird von der International Labor Organisation (ILO) erstellt und unterhalten. Seit 1957 wurden vier Versionen dieser Klassifikation herausgegeben (ISCO-58, ISCO-68, ISCO-88; ISCO-08). Erhebungsprogramme wie z. B. ESS, EVS, ISSP, PIAAC und PISA setzen den ISCO zur Standardisierung von Berufen ein.

  Das Klassifikationsschema ISCO-88 bildet auch den Ausgangspunkt zur Entwicklung der Status- und Prestigemaße

  - des International Socio-Economic Index of Occupational Status (ISEI),
  - der Standard International Occupational Prestige Scale (SIOPS),
  - des Klassenschemas von Erikson, Goldthorpe, Portocarero (EGP) und
  - der European Socio-economic Classification (ESeC).

  Die Maße werden beispielsweise in der European Values Study 2008 angewendet (EVS 2008b. Variable Report – Integrated dataset: 33).
- **KldB** steht für Klassifikation der Berufe in Deutschland. Die aktuelle Version KldB (2010) weist eine hohe Übereinstimmung mit dem ISCO-08 auf. Sie wurde von der Bundesagentur für Arbeit (BA) und dem Institut für Arbeitsmarkt- und Berufsforschung (IAB) federführend entwickelt und systematisch dokumentiert (vgl. Paulus, Schweitzer, Wiemer 2010). Die neue Version hebt die Situation parallel existierender Berufsklassifikationen des Statistischen Bundesamtes und der Bundesagentur für Arbeit auf.

Nationale und internationale Normen geographischer und sprachlicher Einheiten werden bei der Planung von Umfragen zur standardisierten Codierung entsprechender Kontextinformationen genutzt, um die Vergleichbarkeit von Daten zu erleichtern.

- **ISO 639** ist die internationale Norm zur Codierung von Namen für Sprachen. Sie besteht aus insgesamt sechs Teilnormen, die seit 1998 schrittweise eingeführt wurden.

  Die Codetabellen der Teilnorm ISO-639-3 aus dem Jahr 2007, die alle Einzelsprachen und deren Dialekte mit drei Buchstaben codiert, wird von der Nichtregierungsorganisation „SIL International" gepflegt und bereitgestellt. Die Codierung von Sprachen wird im Zusammenhang mit der Dokumentation von nationalen Feldfragebögen mit mehreren Sprachversionen in international vergleichen Erhebungen und zur Kennzeichnung entsprechender Dateien in Datenkatalogen eingesetzt.
- **ISO 3166** dient der Codierung von bestehenden Staaten (ISO 3166-1), staatlichen Untereinheiten (ISO 3166-2) und ehemalige Staaten (ISO 3166-3).

Diese Informationen werden u. a. zur Kennzeichnung der Nationalität in demographischen Angaben oder von Fragebögen und länderspezifischen Datensätzen z. B. in integrierten Datenfiles genutzt. ISO stellt die Listen zu ISO 3166-1 und ISO 3166-2 kostenfrei für den internen und nichtkommerziellen Gebrauch in Englisch und Französisch bereit. Ein Newsletter informiert über Aktualisierungen dieser Codes.

Bei der Dokumentation von integrierten Datensätzen aus unterschiedlichen Ländern werden bestehende bzw. ehemalige Staaten entsprechend ISO 3166-1 und ISO 3166-3 codiert (vgl. z. B. EVS 2008a: 24). Datenkataloge nutzen Ländercodierungen z. T. gemeinsam mit ISO 639-1 zum systematischen Management und zur Erschließung entsprechender Dateien und Studienbestände.

- **NUTS**, die hierarchische Systematik der europäischen Gebietseinheiten für die Statistik, wird von EUROSTAT, dem europäischen Amt für Statistik, herausgegeben. Sie stellt eine Nomenklatur zur systematischen Klassifikation und Identifikation von räumlichen Einheiten in den Mitgliedsstaaten der Europäischen Union dar. Die Systematik wird zur Vergleichbarkeit in der amtlichen Statistik sowie für sozioökonomische Analysen von Regionen und der europäischen Förderpolitik von Regionen angewendet.

Das Schema ist jeweils für drei Jahre gültig. Die aktuelle NUTS-Systematik gilt für den Zeitraum 2012-2014. Nationale Änderungen regionaler Einheiten werden nach drei Jahren in eine neue Systematik aufgenommen. Die hierarchische Struktur wird auf Grundlage von Verwaltungseinheiten gebildet.

Auf Ebene der Nationalstaaten (gemäß ISO-3166-1) werden drei Einheiten unterschieden. In Deutschland umfasst NUTS 1 die Bundesländer, NUTS 2 u. a. Regierungsbezirke und NUTS 3 (Land-)Kreise und kreisfreie Städte.

Internationale Umfragen im europäischen Raum wie z. B. die Eurobarometer oder die European Values Study codieren Variablen zur Region anhand von NUTS Codes.

- Die **Regionalen Standards** (2005) beschreiben verschiedene Instrumente, Techniken und Daten, um regionale Kontexte der Bundesrepublik Deutschland in Vergleichen von Forschungsdaten auf unterschiedlichen regionalen Ebenen einbeziehen zu können.

Die Regionalen Standards und die Demographischen Standards (s. o.) werden als gemeinsame Empfehlungen des Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute e. V. (ADM), der Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V. (ASI) und des Statistischen Bundesamtes (destatis) herausgegeben.

- **ISO 19115:2003** ist der Standard zur Dokumentation geographischer Information und damit verbundener Dienstleistungen. Die Beschreibung von Geodaten durch Metadaten umfasst u. a. Informationen über den Inhalt, Qualität, raum-zeitliche Kontexte der Daten und die Nutzung bzw. Bereitstellung dieser Daten. Eine deutsche Übersetzung der Metadatenfelder aus dem Annex B des Standards ISO 19115 wird von der Geodateninfrastruktur Deutschland (GDI-DE 2008) bereitgestellt.

4.2 Der DDI-Standard zur Dokumentation sozialwissenschaftlicher Studien

Die Data Documentation Initiative (DDI) entstand aus einem 1995 vom amerikanischen Datenarchiv ICPSR initiierten Projekt zur Verbesserung der Möglichkeiten einer standardisierten Dokumentation sozialwissenschaftlicher Daten. In das Vorhaben flossen auch zahlreiche Überlegungen der sozialwissenschaftlichen Datenarchive ein, die zu dieser Zeit bereits international stark vernetzt waren.

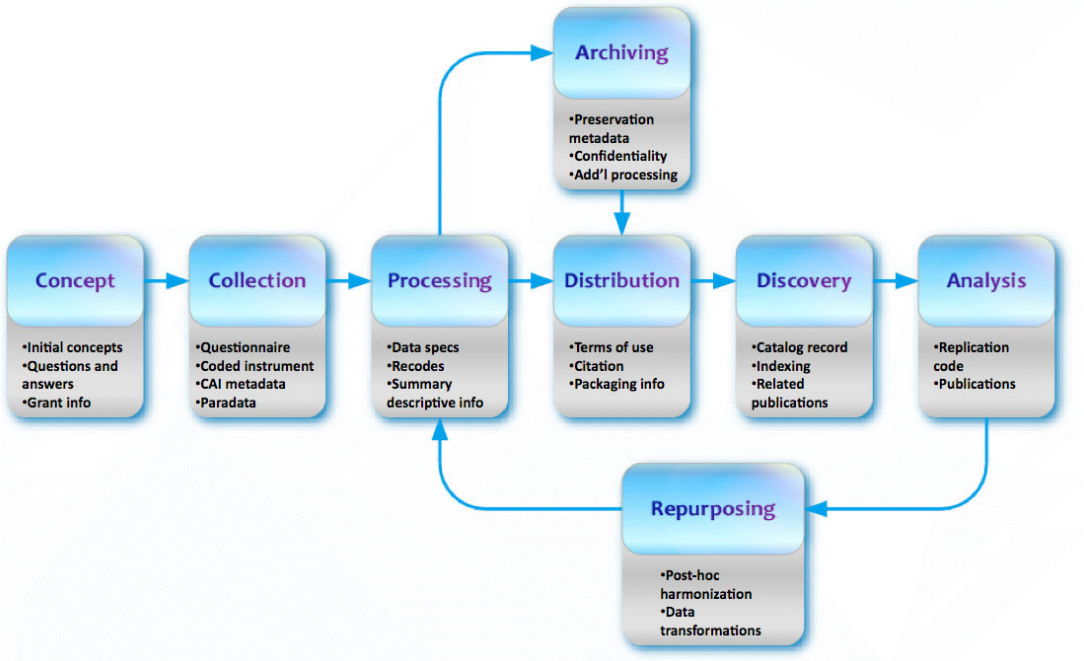
Mit der Gründung der DDI-Alliance im Jahr 2003 wird die Entwicklung von Standards und Instrumenten zur formalen Beschreibung sozialwissenschaftlicher Daten systematisch betrieben. DDI wird durch die Mitgliedschaft von Archiven, Universitäten, Datenzentren, Erhebungsinstituten, Umfrageprogrammen und statistischen Ämtern organisatorisch getragen.

Die aktuellen DDI-Spezifikationen sind in zwei unterschiedlichen Entwicklungslinien von Metadatenstrukturen aufgeteilt:

- DDI-Codebook bzw. DDI-C (bis Version DDI 2; erstmals 2003 veröffentlicht) und
- DDI-Lifecycle bzw. DDI-L (ehemals DDI 3, erstmals 2008 veröffentlicht).

Die Alliance verfolgt in den letzten Jahren mit der Entwicklung der DDI-Lifecycle Version das Ziel, den Lebenszyklus sozialwissenschaftlicher Forschungsdaten möglichst vollständig durch Metadaten zu beschreiben. Die wesentlichen Strukturen und Informationen, die diese zwei Standards kennzeichnen, werden im Folgenden kurz vorgestellt.

Abbildung 3: Unterstützung des Forschungsdatenzyklus durch die DDI-Lifecycle Spezifikation



(Quelle: DDI-Alliance)

4.2.1 Die DDI-Codebook (DDI-C) Spezifikation

Die DDI-Codebook Spezifikation dient der Dokumentation einfacher Umfragen und erlaubt die Beschreibung von Mikrodaten, Aggregatdaten und geographischen Angaben. Die wesentlichen Gruppen von Metadaten und -elementen der Version 2.1 werden in der folgenden Übersicht auszugsweise vorgestellt. Die aktuelle Version 2.5 erleichtert im Wesentlichen die Migration nach DDI-L (DDI3.x) durch die Ergänzung notwendiger Elemente. Die umfassend dokumentierte Spezifikation wird im Web bereitgestellt (siehe Anhang A.4).

Übersicht 9: DDI-Codebook Spezifikation (v.2.1) - Metadatenstruktur und -elemente (Auszug)

Gruppe	Elemente (Auszug)
Document description	▪ Titel, Autoren, Publikation, Zitation des DDI Dokuments
Study description	▪ Primärforscher, Institutionen, Projektinformationen ▪ Typ der Studie, Art der Daten und Abstract ihrer Inhalte ▪ Zeitliche und geographische Angaben zur Erhebung ▪ Methoden, Grundgesamtheit, Analyseeinheit, Stichprobe ▪ Literatur des Projektes zur Studie (z. B. Forschungsbericht; Analysen) ▪ Versionierung, bibliographische Zitation der Datendatei ▪ z. B. Verweise auf digitalisierte Fragebogen, Methodenberichte o. ä.
Variable description	▪ Variablennamen, Typ und Labels ▪ Code, Typ und Häufigkeiten ▪ Hinweise zur Codierung oder Berechnung von Werten (Notes) ▪ Fragetexte und Antwortkategorien ▪ Intervieweranweisungen, Filter
Data files description	▪ Anzahl der Variablen, Fälle ▪ Namen, Formate und Versionen

4.2.2 Die DDI-Lifecycle (DDI-L) Spezifikation

Die DDI-Lifecycle Spezifikation (DDI-L, aktuelle Version 3.1) basiert auf einer umfangreichen modularen Erweiterung des codebuch- und variablenzentrierten Metadatenansatzes, um u. a.

- den gesamten Lebenszyklus von Forschungsdaten mit Metadaten zu beschreiben und das Datenmanagement in Forschungsprojekten, Archiven und anderen beteiligten Institutionen zu unterstützen;
- einfache und komplexe Daten möglichst vollständig und genau durch vielfältige Metadaten zu beschreiben und dadurch das Wissen über das Analysepotential der Daten für Nutzer zu transportieren;
- einmal erfasste Metadaten für vielfältige Zwecke und Anwendungen wieder verwendbar zu machen, z. B. bei der Replikationen von Erhebungen in der Datendokumentation, zur Daten-

definition für verschiedene Statistikprogramme, bei differenzierten Recherchen in Daten- und Fragenbeständen u. a. m.

Durch den modularen Aufbau ist es möglich, Metadaten bereits dort zu erfassen, wo sie entstehen und mit weiteren Kooperationspartnern auszutauschen bzw. im Rahmen der Arbeitsteilung weiterzuleiten. So unterstützt DDI-L die Erstellung von Studien- und Datendokumentationen und deren Erschließung in den Phasen

- der konzeptuellen Entwicklung des Studiendesigns,
- des Untersuchungsdesigns und der Datenerhebung,
- der Datenaufbereitung,
- der Datenarchivierung und Datenbereitstellung,
- der Datenerschließung und Sekundärnutzung von Daten.

Die Analyse vorhandener Daten kann von Detailkorrekturen an den Daten (Repurposing / Processing) bis hin zur Durchführung neuer Untersuchungen reichen, wodurch die entsprechenden Phasen im Lebenszyklus von Forschungsdaten neu durchlaufen werden.

Wesentliche Module und zugehörige Typen von Dokumentationselementen der DDI-Lifecycle Spezifikation Version 3.1 sind auszugsweise in der folgenden Tabelle dargestellt.

Übersicht 10: DDI-Lifecycle Spezifikation - Module und Dokumentationselemente (Auszug)

Modul	Dokumentationselemente (Metadaten)
Study Unit	▪ Umfangreiche Metadaten zur Identifikation einer Studie (Titel, Forscher), zur bibliographischen Zitation und zur Nutzung dauerhafter Identifikatoren z. B. in Form einer DOI (Digital Object Identifier) ▪ Abstract, Zweck, Antrag und Förderung einer Einzelstudie ▪ Thematische, zeitliche und räumliche Abdeckung der Studie ▪ Untersuchungseinheit(en), Messkonzepte der Daten ▪ Publikationen und Verweise auf digitalisierte Materialien: Fragebogen, Methodenberichte o. ä.
Data Collection	▪ Methodologie; Erhebungszeiträume und -ereignisse ▪ Mehrsprachige Fragetexte und Antwortdomänen (Text, Numerisch, Codeschema etc.) ▪ Messinstrument, Ablauf- und Kontrollstrukturen, Intervieweranweisungen, Codierungsanweisungen
Logical Product	▪ Logische Beschreibung der Forschungsdaten einer Studie. ▪ Variablenschema, Variablendefinition, Kategorien-, Codelisten ▪ NCubes; Gruppierung von NCubes und Variablen zur Beschreibung aggregierter Daten ▪ Logische Beziehungen der Daten untereinander, z. B. durch Kennzeichnung von mehreren Datensätzen (Records) je Befragtem

Modul	Dokumentationselemente (Metadaten)
Physical Data Product Physical Data Structure Physical Instance (> Dataset / NCubes)	▪ Beschreibung der physischen Eigenschaften von (komplexen) Datensätzen und ▪ ihren strukturellen Beziehungen (z. B. rechteckig, hierarchisch, relational) bzw. ▪ der konkreten Datendatei (Variablen, Namen, Formate, Version) und Speicherinformationen (Ort, Fingerprint etc.).
Comparative	Modul zum paarweisen Vergleich von Objekten, von Metadaten mit externen Standards oder Vergleiche innerhalb von Metadatenübersichten hinsichtlich Untersuchungseinheit, Messkonzepten, Fragen, Variablen, Kategorien und Codes.
Archive	Ort an dem sich ein Set von Metadaten zu einem bestimmten Zeitpunkt entlang des Lebenszyklus befindet und in Verantwortung einer sog. „DDI Agency" verantwortlich bearbeitet, gepflegt und / oder aufbewahrt wird. ▪ Organisationsschema: z. B. DDI Agency Identifier, Namen, Rollen, Kontakte einer Organisation oder einer Person ▪ Archivspezifische (technische, administrative, inhaltliche) Informationen im Zusammenhang mit der Datenhaltung (z. B. Bestand, Speicherort einer Studie / Studienkollektionen bzw. Datendateien) und Bereitstellung (Verfügbarkeit, Zugangsregelung, Beschränkungen etc.).

DDC-L unterstützt den Dublin Core Standards (ISO 15836:2009) und setzt weitere Module für spezielle technische bzw. administrative Zwecke ein („Reusable", „DDI Profile").

Die Spezifikation der DDI-Version 3.1 wird im Webangebot der DDI-Alliance umfassend dokumentiert und zusätzlich durch Anwendungsbeispiele in Projekten sowie Use-Cases und Best Practice Empfehlungen ergänzt (vgl. die Quellen zu DDI in Anhang A.4).

Neben der Erweiterung des Umfangs an Metadatentypen und -elementen war ein wesentliches Motiv der Entwicklung des DDI-Lifecycle Standards, die technischen Voraussetzungen so zu verbessern, dass die intensive Wiederverwendung von Metadaten in datenhaltenden Institutionen sowie datenerhebenden Einrichtungen und Umfrageprogrammen auf breiter Basis möglich ist.

Dies wird u. a. durch die sog. „Scheme" basierten Module erreicht. Jedes Schema stellt einen speziellen Objekttyp dar, z. B. Fragen, Kategorien, Variablen, Konzepte. Schemata ermöglichen es,

- Listen mit gleichen Items zu erstellen und zu pflegen,
- Listen zu versionieren und
- Elemente durch Referenzen in Listen wieder zu verwenden.

Die Möglichkeiten der Versionierung und Referenzierung von DDI Metadaten erhöhen die Wiederverwendbarkeit vorhandener Datendokumentationen und erweitert zugleich die Informationsbasis für interessierte Nutzer bei der Erschließung vorhandener Forschungsdaten, u. a. durch die

- hohe Transparenz der Daten durch gezielte Erschließung von Daten in Datenkatalogen bis auf Variablenebene,

- Versionsangabe des Datenfiles und Informationen über Errata,
- Bereitstellung vielfältiger Kontextinformationen (z. B. Literaturhinweise, Dokumente zur Studie etc.) und
- hohe Sichtbarkeit von Daten und Datenproduzenten durch die Möglichkeit der einfachen Zitation der Daten.

Weiterhin werden kontrollierte Vokabulare zu verschiedensten DDI Metadaten entwickelt, um die einheitliche Dokumentation und Erschließung von Studien und Daten zu vereinfachen. Die kontrollierten Vokabulare sind nicht Teil der DDI-Spezifikationen, sondern werden als eigenständige Empfehlung der DDI-Alliance herausgegeben.

Die DDI-Spezifikationen beruhen ab DDI Version 2.5 auf XML (Extensible Mark-up Language), was erheblich zur Vereinfachung der Programmierung von Werkzeugen und der informationstechnologischen Verarbeitung und Erschließung der Informationen beiträgt. Gleichzeitig erlaubt das XML-Format einen nahtlosen Austausch von Metadaten zwischen verschiedenen Organisationseinheiten entlang des Data Lifecycle. Informationen über entsprechende technische Werkzeuge, DDI-kompatible Software und Konvertierungsroutinen (z. B. Transfer von Metadaten aus unterschiedlicher statistischer Software nach DDI) sind ebenfalls auf den Internetseiten der DDI-Alliance verfügbar.

4.3 Persistent Identifier zur dauerhaften Zitation von Forschungsdaten

Dauerhafte Identifikatoren werden schon seit langem genutzt, um etwa gedruckte Bücher oder Zeitschriften durch eine Internationale Standard-Buchnummer (ISBN) eindeutig zu kennzeichnen. Im Bereich digitaler Textpublikationen haben sich darüber hinaus spezielle Persistent Identifier (PI) Systeme (Schroeder 2010) etabliert, um eine eindeutige und dauerhafte Referenzierung und Onlineverfügbarkeit von wissenschaftlichen Artikeln oder anderen Inhalten zu ermöglichen.

Solche PI-Systeme gewinnen in entsprechend modifizierter Form eine wachsende Bedeutung, um bestehende Barrieren bei der systematischen Zitation von Forschungsdaten in wissenschaftlichen Veröffentlichungen zu überwinden. Gleichzeitig verbessert die Nutzung dauerhafter Identifikatoren die Auffindbarkeit und Zugänglichkeit von publizierten Forschungsdaten.

So nutzen etwa die Mitglieder von DataCite das DOI®-System (Digital Object Identifier), um mit Hilfe dieser Technik Forschungsdaten als eigenständige und zitierfähige wissenschaftliche Objekte zu etablieren.

Die wissenschaftsrelevanten Vorteile einer Anwendung von PIs z. B. auf sozialwissenschaftliche Forschungsdaten werden durch folgende Aspekte greifbar:

- kompakte Zitation und dauerhafte Auffindbarkeit von Forschungsdaten;
- erhöhte Sichtbarkeit von Forschungsdaten und Datenanbietern;
- Nachweismöglichkeit des Impacts durch Datenzitationsraten;
- Anerkennung der Datenerstellung und Dokumentation als wertvolle Forschungsleistung;
- verbesserte Nachvollziehbarkeit und intersubjektive Überprüfbarkeit der Forschungsergebnisse;
- Unterstützung des Data Sharings und damit Verbreiterung der Datenbasis für Sekundäranalysen;
- Möglichkeit der Verknüpfung von datenbasierten Forschungspublikationen und zugrundeliegenden Daten über entsprechende Publikationsportale.

4.3.1 DataCite und die Vergabe von DOI-Namen

DataCite ist ein 2009 gegründetes internationales Konsortium von führenden Forschungsbibliotheken und Informationszentren. Es verfolgt das Ziel, einheitliche Standards zur Akzeptanz von Forschungsdaten als eigenständige, zitierfähige wissenschaftliche Objekte weltweit zu fördern und zu etablieren. Datenserviceeinrichtungen können auf Grundlage des DOI®-Systems ihre Forschungsdaten mit Hilfe von DOI-Namen registrieren lassen, damit sie dauerhaft zitierbar und zugänglich vorgehalten werden können.

DataCite ist bei der International DOI Foundation (IDF) als Registrierungsagentur akkreditiert. Die IDF verwaltet und entwickelt das DOI®-System. Deutsche Mitglieder in DataCite sind

- die Technische Informationsbibliothek Hannover (TIB),
- die Deutsche Zentralbibliothek für Medizin (ZB MED),
- die Deutsche Zentralbibliothek für Wirtschaftswissenschaften (ZBW) und
- das Leibniz-Institut für Sozialwissenschaften (GESIS).

Sie vergeben in der Funktion einer DOI Allocation Agency die DOI®-Namen für Forschungsdaten der jeweiligen Domäne an entsprechende Datenanbieter (sog. Publikationsagenten). Die TIB Hannover als Managing Agent von DataCite organisiert die Präfixverwaltung und die Verbindung zur IDF.

Funktion, Struktur und Darstellung einer DOI® (Digital Object Identifier)

DOI-Namen sind eine von mehreren technischen PI Lösungen, um physische oder digitale Objekte (u. a. Daten, Dokumente, Publikationen, andere Objekte) dauerhaft zu referenzieren und Metadaten stabil mit diesem Objekt zu verknüpfen. Dazu wird besonders betont, das DOI-Namen als „"digital identifier of an object" (not "identifier of a digital object")" konstruiert werden (IDF: □DOI®Handbook2012).

Die DOI verweist auf das jeweilige Objekt oder eine entsprechende Onlinebeschreibung (landing page) des Anbieters. D. h., ein DOI-Name wird dauerhaft mit dem Objekt und nicht mit dem spezifischen Ort der Speicherung verknüpft. In einem weiteren Schritt wird dann durch eine URL, die über das zugehörige Metadatenschema eingepflegt wird, auf den jeweils aktuellen Standort des Objektes auf einem Server verwiesen. Dadurch lassen sich Defizite der unmittelbaren Adressierung von Online-Ressourcen durch URLs ausgleichen (Fehlermeldung wegen Abschaltung, Unerreichbarkeit, Umzug oder Reorganisation des Servers).

Eine zentrale Voraussetzung für die Qualitätssicherung eines solchen Systems ist, dass die technische Infrastruktur und die organisatorischen Maßnahmen den Zugang und die Aktualität der Ressourcen langfristig gewährleisten. So bleiben zitierte Forschungsdaten auch dann über deren DOI-Namen online erreichbar, auch wenn z. B. der Speicherort vom Datenanbieter intern verlegt wird. Es müssen vor Verlegung des Speicherortes die entsprechenden Metadaten in der Datenbank der DOI-Registrierungsagentur so angepasst werden, dass der mit einem DOI-Namen verbundene Link (URL) auf den neuen Standort zeigt. Die mit dem DOI-Namen verknüpften Metadaten beschreiben z. B. einen Datensatz durch bibliographische, inhaltliche und methodische Informationen (siehe 0).

Ein DOI-Name ist eine weltweit eindeutige alphanumerische Zeichenfolge. Ein DOI-Name besteht aus zwei Teilen: dem Präfix und dem durch einen Slash getrennten Suffix, z. B. 10.4232/1.10834 (vgl. die nächste Abbildung).

Abbildung 4: Struktur eines DOI-Namens

Die „10" identifiziert den Typ des gesamten Identifiers (Präfix und Suffix) als DOI. Der zweite Teil des Präfixes ist der eindeutige (ab 1000 aufsteigende) Code einer registrierten Organisation (Datenzentrum bzw. Datenanbieter), die Objekte (Dokumente, Forschungsdaten etc.) im DOI-System eingetragen hat. Dieser Code des sog. Publikationsagenten bleibt dauerhaft gültig, unabhängig von späteren Änderungen der Rolle dieser Einrichtung bei der Unterhaltung der angemeldeten DOI-Namen. Das Präfix wird von der IDF verwaltet und durch eine Registrierungsagentur wie DataCite und den dazugehörigen Allocation Agencies den Publikationsagenten zugewiesen.
Das Suffix kennzeichnet das registrierte Objekt. Es kann entweder frei oder durch die Anwendung verschiedener Schemata systematisch gebildet werden (vgl. IDF: □DOI®Handbook 2012). Dabei ist der Empfehlung vom IDF zu folgen, keine sprechenden Zeichenfolgen zu verwenden.

Publikationsagenten müssen eine Policy für die Versionierung ihrer Forschungsdaten festlegen. Dabei wird festgelegt, welche Änderungen an Daten/Metadaten zur Registrierung eines neuen DOI-Namens führen. Für Forschungsdaten, die durch die datenhaltende Einrichtung versioniert werden, muss jeweils ein neues Suffix erzeugt werden. D. h., eine neue Version eines Forschungsdatensatzes wird erneut registriert und erhält damit einen eigenen DOI-Namen (vgl. auch „Versionierung von Datensätzen" Abschnitt 3.2, .S. 41).

Namen von Digital Object Identifiern können je nach Nutzung, z. B. im Zusammenhang mit der vollständigen Zitation eines Datensatzes (vgl. Übersicht 11: S. 55), auf unterschiedliche Weise dargestellt werden:

- In (Off-line) Medien kann ein DOI-Name durch ein vorangestelltes, kleingeschriebenes „doi:" (analog zu Protokollen wie „http:" oder „ftp:") gezielt gekennzeichnet werden, z. B. „doi:10.4232/1.10834".

  Allerdings führt ein entsprechender Eintrag in das Adressfeld des Browsers nicht ohne weiteres zur Online-Quelle, da Browser ohne ein zusätzliches Plug-in das Protokoll „doi:" zurzeit meistens nicht erkennt. Die Adresse wird nur richtig „aufgelöst", wenn der DOI-Namen in das Textfeld eines sog. DOI Resolver Dienstes (z. B. DataCite) eingegeben wird.

- Um die Quelle des Objektes direkt ansprechen zu können, sollte die DOI deshalb entweder mit der URL des Resolvers abgedruckt „http://dx.doi.org/doi:10.4232/1.10834" oder platzsparender mit einem Hyperlink „doi:10.4232/1.10106" unterlegt werden.

Derzeit werden von da|ra, der Registrierungsagentur für Sozial- und Wirtschaftsdaten, zwei Vorschläge für Zitationen unterbreitet, wobei in beiden Zitationsweisen der Identifier optional im Original- oder in einem http-Format erscheinen kann (vgl. Übersicht 11: S. 55):

1. Primärforscher (Veröffentlichungsdatum): Titel. Publikationsagent. Identifier (vgl. Beispiel 1)
2. Primärforscher (Veröffentlichungsdatum): Titel. Version. Publikationsagent. Typ der Ressource. Identifier (vgl. Beispiel 2)

Übersicht 11: Zitation von Forschungsdaten mittels eines DOI-Namens

Beispiel 1

Mansel, Jürgen (2009): Psychosoziale Belastungen Jugendlicher. Primärdaten einer Längsschnittstudie zum Erleben des Golfkrieges. ZPID - Leibniz-Zentrum für Psychologische Information und Dokumentation. doi:10.5160/psychdata.mljn91be12

Beispiel 2

ISSP Research Group (2012): International Social Survey Programme: Environment III - ISSP 2010. Version 2.0.0. GESIS Data Archive, Cologne. ZA5500 Data file. doi:10.4232/1.11418

4.3.2 da|ra - Registrierungsagentur für Sozial- und Wirtschaftsdaten

Die GESIS Leibniz-Institut für Sozialwissenschaften und die ZBW Leibniz-Informationszentrum Wirtschaft betreiben in Kooperation mit DataCite auf Grundlage des DOI-Systems die nicht-kommerzielle Registrierungsagentur für Sozial- und Wirtschaftsdaten da|ra (www.da-ra.de).

Die Registrierungsagentur startete 2010 mit einem Pilotprojekt, um zunächst ein technisches und organisatorisches Konzept zu erarbeiten.

Mitte 2010 ging das Angebot für sozialwissenschaftliche Daten online und enthält mittlerweile über 5500 registrierte Studien und mehr als 8000 Metadatensätze – hierunter ca. 2400 Metadatensätze der OECD iLibrary, die über die DOI-Registrierungsagentur crossref in das Informationssystem übernommen wurden (Hausstein 2012: 10).

Über den da|ra Service können Datenanbieter Umfragedaten, Aggregatdaten, Mikrodaten und qualitative Daten registrieren lassen. Mit der Übermittlung der erforderlichen Metadaten erhält jeder Datensatz einen eindeutigen DOI-Namen.

da|ra bezieht die DOI-Namen über die GESIS Mitgliedschaft in DataCite. Die Vergabe des Suffixes für den Datenfile erfolgt durch den Publikationsagenten (Datenanbieter) zusammen mit da|ra nach der Registrierung.

Die Vergabe von DOI-Namen wird in der da|ra Policy beschrieben. So werden u. a. die rechtlichen Voraussetzungen und Rollen der Beteiligten im Registrierungsprozess festgelegt. Sie benennt weiterhin die technischen Anforderungen an die zu registrierenden Objekte und an die Speichersysteme auf Seiten des Datenanbieters.

Bestandteil dieser Policy ist der Abschluss eines bilateralen Service Level Agreement zwischen da|ra und dem jeweiligen Publikationsagenten.

Mit dem Abschluss dieser Vereinbarung werden Arbeitsabläufe und Verfahren bei der Datenregistrierung sowie die Verantwortlichkeiten der Beteiligten schriftlich geregelt. Dazu zählen u. a.

- die dauerhafte Zugänglichkeit zu den Daten durch entsprechende technische Dienste und Speichersysteme,
- die Bereitstellung und Aktualität der Metadaten,
- die Versionierung der Objekte sowie
- Belange des Urheberrechtes und der Veränderungen der Datenpublikationsaktivität auf Seiten des Publikationsagenten (vgl. da|ra 2012).

da|ra Metadatenschema für sozialwissenschaftliche Studien und Daten

Im Verlauf des Aufbaues dieses Dienstes wurde das Metadatenschema von DataCite (2011) um spezifisch sozialwissenschaftliche Metadaten auf Grundlage des GESIS Datenbestandskataloges (Zenk-Möltgen, Habbel 2012) ergänzt.

Das zunächst in der Version 1.0 verwendete Metadatenschema ist mit dem Metadatenschema des DDI Standards kompatibel und gewährleistet zusammen mit der Anwendung kontrollierter Vokabulare die Interoperabilität der registrierten Datenquellen mit anderen Datenbeständen auf internationaler Ebene.

Um die internationale Erschließung und Präsentation der Daten zu sichern, ist auch eine englische Version der obligatorischen und optionalen Metadatenelemente erforderlich. Das aktuelle da|ra Schema in der Version 2.2.1 wurde erweitert, um kompatibel zum aktuellen DataCite Metadatenschema zu bleiben (vgl. Hausstein et al. 2012).

Die Metadaten umfassen neben den für die Zitation wichtigen Elementen zusätzliche inhaltliche und formale Beschreibungen des registrierten Forschungsdatensatzes, der Datenerhebung und der beteiligten Personen und Institutionen.

Das Schema besteht aus 3 Gruppen von Informationen: Pflichtelemente, optionale und administrative Elemente. Einzelne Metadatenelemente werden mit Hilfe von domänenspezifischen Klassifikationen, Thesauri oder kontrollierten Vokabularen beispielsweise für Personen oder Institutionen (Personennormdateien u. Ä.) definiert und erfasst.

5 Von der Sicherung zur langfristigen Nutzung der Forschungsdaten

Zum Projektende stellen sich eine Reihe von Fragen zum weiteren Umgang mit den erhobenen Daten und den vom Forschungsprojekt erstellten Materialien und Dateien (Erhebungsinstrumente, Codepläne, Feldbericht, Datenfiles, Syntaxdateien etc.).

Die abschließend durchgeführten Maßnahmen zur Datensicherung sollten projektintern dokumentiert werden. Dies ist unabhängig davon, ob die Daten etwa für zehn Jahre an einem Forschungsinstitut aufbewahrt werden oder einer Einrichtung zur institutionellen Langfristarchivierung übergeben werden.

Der folgende Abschnitt 5.1 skizziert zunächst mögliche organisatorische Optionen und Fragestellungen zur Datensicherung und -bereitstellung, die bereits in die Projektplanung einfließen können.

Anforderungen an die Dokumentationen der Daten und ihren Entstehungszusammenhang aus der Sicht der Qualitätssicherung und längerfristigen Nutzbarkeit der Daten werden anschließend zusammen mit standardisierten und DDI konformen Metadaten auf Studien und Variablenebene thematisiert (5.2).

Allgemeine rechtliche Fragen der Langzeitsicherung und die vertragliche Regelung der Langzeitarchivierung im GESIS Datenarchiv werden in gesonderten Abschnitten (5.3; 5.4) behandelt. Die Auswahl und Übergabe der zu sichernden Daten und Dokumentationen wird abschließend angesprochen (5.5).

5.1 Optionen von der Datensicherung bis langfristigen Archivierung von Daten

Eine praktische, zu Projektbeginn zu klärende Frage für ein datenerzeugendes Projekt ist, welche Daten und Studienmaterialien, wo, wie lange und an welchem Ort nach Projektabschluss gesichert bzw. langfristig archiviert werden können. Daran schließt sich die Frage an, wer unter welchen Bedingungen Zugang zu den so gesicherten Projektergebnissen haben soll. Übersicht 12 gibt einen Überblick über verschiedene organisatorische Optionen zur Datenaufbewahrung nach Projektende.

Zur „Sicherung guter wissenschaftliche Praxis" empfiehlt die DFG seit 1998: „Primärdaten als Grundlagen für Veröffentlichungen sollen auf haltbaren und gesicherten Trägern in der Institution, wo sie entstanden sind, für zehn Jahre aufbewahrt werden." (DFG 1998:12f).

Aktuell sind von der DFG geförderte Projekte gehalten, „Daten, die für Nachnutzung geeignet sind", „nachhaltig zu sichern und ggf. für eine erneute Nutzung bereitzustellen". Bei der Festlegung von Maßnahmen der Datensicherung und möglichen Nachnutzung sollen existierende fachspezifische Standards und Angebote datenhaltender Einrichtungen berücksichtigt werden (DFG 2012c).

Mit Blick auf die Ermöglichung einer Sekundärnutzung von Primärdaten aus einem Forschungsprojekt ist eine zehnjährige Aufbewahrung sicherlich ein erster Schritt.

Eine nachhaltige Nutzung von Daten erfordert jedoch weitergehende Maßnahmen, die die Archivierung und Bereitstellung technisch und organisatorisch auch langfristig ermöglichen. Entsprechende Beratungsmöglichkeiten und Archivierungsangebote von Datenserviceeinrichtungen sollten deshalb wahrgenommen werden (vgl. Übersicht 13).

Übersicht 12: Organisatorische Optionen von der Datensicherung bis zur Langzeitarchivierung

- Speicherung in der Abteilung oder am Institut zur zehnjährigen Sicherung der Daten;
- Sicherung und/oder Bereitstellung der Daten durch einen universitären Datenservice;
- Einrichtung eines projekteigenen Datenrepositoriums zur Sicherung und/oder Bereitstellung der Forschungsdaten (i. d. R. nur bei Großprojekten);
- Langfristige Archivierung und Bereitstellung durch ein disziplinspezifisches Forschungsdatenzentrum oder ein domänenspezifisches Datenarchiv wie das GESIS Datenarchiv für Sozialwissenschaften (vgl. Kapitel 6; Anhang A.1).

Neben der Wahl einer organisatorischen Lösung sind weitere Aspekte wie die zeitliche Dauer der Datensicherung und Fragen der Datennutzung zu berücksichtigen.

Übersicht 13: Leitfragen zur Datensicherung und Langzeitarchivierung

- Daten, Datendokumentationen und weitere studienrelevante Materialien:
  Welche Daten, Projektdokumente und Methodeninformationen sind aufzubewahren?
- Datenhaltung und Aufbewahrungsfrist:
  Sollen sie 10 Jahre oder länger aufbewahrt werden? Garantiert die gewählte Organisationsform oder die Institution technisch und administrativ die Nutzbarkeit der Daten und Dokumentationen im angestrebten Zeithorizont? Entstehen urheberrechtliche und datenschutzrelevante Fragen für die datenhaltende Einrichtung und wie werden diese geregelt?
- Datenbereitstellung:
  Sollen bzw. können die Forschungsdaten aus Sicht des Projektes bereitgestellt werden? Welche technischen Dienste sind verfügbar, um die Forschungsdaten und Dokumentationen zu erschließen und bereitzustellen?
- Regelungen zum Datenzugang:
  Unter welchen Zugangsbedingungen können die Daten für Sekundäranalysen kurzfristig und/oder langfristig bereitgestellt werden? Kann der Zugang zeitverzögert erteilt werden („Embargozeit", „moving wall"), z. B. zur Sicherung von Erstauswertungen durch das Forschungsprojekt?
- Datenschutz:
  Müssen Forschungsdaten vor der Weitergabe anonymisiert werden oder können die Daten nur in einer besonders geschützten Umgebung (Safe Center) genutzt werden?

Die Kooperation mit einer datenhaltenden Einrichtung vereinfacht die Durchführung entsprechender Sicherungsmaßnahmen zum Projektende, wenn inhaltliche und formale Belange bereits abgesprochen wurden.

5.2 Projektdokumentation - Metadaten auf Studien- und Datenebene

Metadaten auf Studienebene dienen u. a. der Kurzbeschreibung der Fragestellung des Forschungsvorhabens, der Projektbeteiligten, des Designs der Studie sowie der eingesetzten Methoden und Messinstrumente. Solche inhaltlichen Beschreibungen, ergänzt durch strukturelle Metadaten zu den Daten, unterstützen im Sinne von Kataloginformationen die nutzerfreundliche Erschließung archivierter Stu-

dien in Datenkatalogen. Sie werden auf Grundlage der studienbezogenen Materialien erstellt, die zusammen mit den Forschungsdaten langfristig gesichert werden. Aus Sicht der nachhaltigen Nutzung der Daten stellen die Datendokumentation, das Messinstrument und der Methodenbericht die drei wesentlichen Elemente dar, die für die Sekundäranalysen archivierter Daten verfügbar sein müssen. Das GESIS Datenarchiv dokumentiert alle damit zusammenhängenden Metadaten auf Studien-, Fragen- und Variablenebene systematisch nach dem DDI-Standard.

5.2.1 Leitfragen zur Beschreibung des Methodendesigns

In der sozialwissenschaftlichen Fachliteratur wird insbesondere auf die methodischen Detailinformationen hingewiesen, die erforderlich sind, um Forschungsdaten durch Sekundäranalysen oder zur Replikation von Ergebnissen wissenschaftlich adäquat nutzen zu können (Schnell, Esser, Hill 2011: 485f; Häder 2010: 449f). Die Facetten des Methodendesigns werden in einem sozialwissenschaftlichen Forschungsvorhaben üblicherweise durch Methodenberichte, Feldberichte oder Technische Reports dokumentiert. Dabei stellt sich für Forscher, die ihre Daten langfristig sichern und bereitstellen wollen, möglicherweise die Frage, welche Methodeninformationen sie konkret einer Datenserviceeinrichtung zur Verfügung stellen. In der Beratung von Datengebern oder Projektvorhaben durch das GESIS Datenarchiv werden diese Aspekte anhand eines Leitfadens zur Erstellung von Methodenberichten für die Datenarchivierung besprochen (Watteler 2010a). Die wesentlichen Fragenstellungen zeigt Tabelle 12.

Übersicht 14: Leitfragen zur Erstellung von Methodeninformationen für die Datenarchivierung

1. Studiendesign (Konzept):
   Beschreibung von Fragestellung, Anlass und Zweck der ursprünglichen Forschung.
   - Was wurde an Daten erhoben? Werden vorhandene Daten ausgewertet?
     Art der Daten, Herkunft existierender Daten, technisches Format.
   - Wer war am Projekt und der Datenerhebung beteiligt?
     Primärforscher, Feldforschungsinstitut, Auftraggeber, Fördereinrichtung, o. ä.
2. Stichprobe (Sampling)
   Beschreibung, wie die Daten erhoben wurden.
   - Grundgesamtheit, Auswahlgesamtheit, Stichprobenverfahren, Modus der Erhebung, Kopie des Messinstrumentes einschließlich aller benutzten Hilfsmittel.
3. Datenerhebung (Pretest; Feldphase)
   Beschreibung wann und wo die Daten erhoben wurden.

- Pretest(s); Ergebnisse, Ereignisse, Kontrollen der Feldarbeit; Informationen über Interviewer, Kontaktversuche, Brutto- / Nettostichprobe; Ort der Befragung, regionale Ausdehnung.

4. Datenaufbereitung
   Beschreibung zu Datensatz, Datenqualität, Datenschutz.
   - Wie wurde der Datensatz erstellt, vercodet und korrigiert?
   - Wie wurden die Daten hinsichtlich Reliabilität, Validität und Repräsentativität geprüft?
   - Wurden die Daten aus Datenschutzgründen ggf. anonymisiert?

5.2.2 Metadaten zur Beschreibung der Studie

Im Zuge einer dauerhaften Datensicherung stellen Metadaten auf Studienebene die wissenschaftlich relevanten Informationen zur Erschließung von Studien und Forschungsdaten bereit. So dokumentieren die Metadaten in einer Studienbeschreibung durch Kurzbeschreibungen oder systematische Vokabulare z. B. über das Stichprobendesign, Verfahren der Stichprobenziehung u.v.m.

Die Metadaten auf Studienebene werden im GESIS Datenarchiv z. B. aus Projekt- und Methodenberichten extrahiert und in einem standardisierten Studienbeschreibungsschema erfasst (Bauske 2000; Zenk-Möltgen, Habbel 2012). Das DDI basierte Schema wird von sozialwissenschaftlichen Datenarchiven weltweit zur Dokumentation und Erschließung archivierter Studien in Datenkatalogen und -portalen eingesetzt. Sie erlauben u. a. die verknüpfte Recherche in den wesentlichen Metadaten auf Studienebene, wie die erweiterte Suche im Datenbestandskatalog der GESIS zeigt (vgl. Anhang A1).

Abbildung 5: Metadatenrecherche im Datenkatalog der GESIS

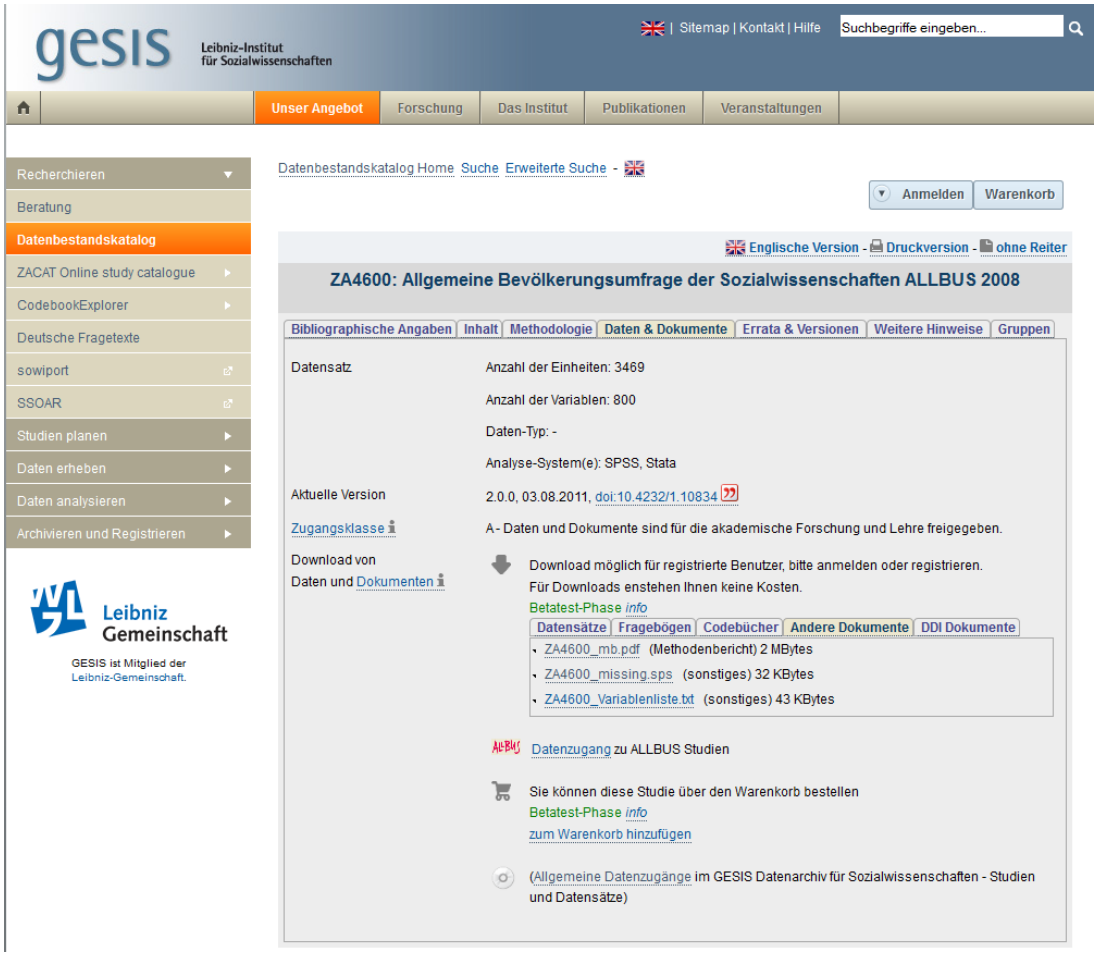
Im Datenbestandskatalog kann in folgenden Metadaten recherchiert werden:

- Titel der Untersuchung, originalsprachlicher Titel der Studie, Studiennummer, Erhebungsjahr,
- Primärforscher / wissenschaftlicher Beirat; Institution, die die Daten erhoben hat,
- Inhalt (Abstract der Studie), Untersuchungsgebiet, Grundgesamtheit; Auswahlverfahren, Erhebungsverfahren, Angaben zum Datensatz und □Analyse□System,
- Veröffentlichungen auf Grundlage der Studie und Hinweise z. B. auf weitere Studien, inhaltliche Kategorien und Zugangsklasse.

Aus der Ergebnisliste können dann eine oder mehrere Studienbeschreibungen ausgewählt werden. Die einzelne Studienbeschreibung enthält umfangreiche Metadaten u. a. zu Bibliographie, Inhalten und Methodologie der Studie sowie zu Versionen der Daten (siehe Abbildung 6). Im aktiven Segment „Daten & Dokumente" informiert beispielsweise die Studienschreibung zum Datensatz ZA4600 (ALLBUS 2008b) über

- den Datensatz, die aktuelle Version und zugehörigen Persistent Identifier (DOI®),
- Downloadmöglichkeiten des Datensatzes (Reiter Datensätze) und
- Dokumentationen zur Studie (Reiter „Fragebögen" bzw. „Codebücher") und den Zugang zum Methodenbericht, Missing Definitionen und Variablenliste (Reiter „Andere Dokumente").

Abbildung 6: Beispiel der Studienbeschreibung ALLBUS 2008 (Ausschnitt)



5.2.3 Metadaten auf Fragen- und Variablenebene

Die Metadaten auf Fragen- und Variablenebene dokumentieren die inhaltliche Bedeutung der Daten auf Grundlage der Fragenbogendesigns und der Definition der Variablen (vgl. Abschnitt 2.1). Diese Informationen werden u. a. durch Einzeldokumente wie den Codeplan, eine Variablenliste sowie durch den Originalfragebogen selbst bereitgestellt.

Eine aufwendigere Dokumentationsform stellt das Codebuch oder der Variablenreport dar, in dem Fragen, Variablen und Randauszählungen zusammenhängend in einem Dokument dokumentiert werden (vgl. Abbildung 7, S. 64).

Die Frage, welche Dokumentations- und Publikationsform im Hinblick auf die projektbezogene Sicherung oder eine langfristige Archivierung der Studie gewählt wird, muss im jeweiligen Projektkontext entschieden werden. Im Folgenden werden relevante Metadaten zum Messinstrument (Fragenbogen) und den Variablen im Datenfile dargestellt, die zur längerfristigen Nutzbarkeit einer sozialwissenschaftlichen Studie verfügbar sein müssen.

Dokumentation der Fragen des Fragebogens sowie der Variablen und Datenmodifikationen

Der standardisierte Fragebogen, sein Aufbau und die Ablaufstruktur der einzelnen Fragen bilden die methodische Grundlage der geplanten Datenerhebung. Mit Blick auf die längerfristige Nutzung der Daten stellt das Messinstrument zusammen mit dem Methodenreport zugleich die wesentlichen Metadaten zum methodischen und inhaltlichen Verstehen der Daten bereit. Im Zuge der PAPI basierten Erhebungstechnik liegen alle Metadaten zumindest in Papierform als Kopie des Fragebogens und von weiteren Hilfsmitteln (Listen, Karten etc.) vor. Im Zuge der Nutzung von computerbasierten Erhebungsverfahren müssen entsprechende elektronische Versionen gesondert erzeugt werden.

Zum dauerhaften Verständnis der Daten müssen alle Elemente des Fragebogens bzw. seine Fragen dokumentiert sein, und auch als maschinenlesbares Dokument zur längerfristigen Sicherung gespeichert werden. Zu den Metadaten zählen folgende Informationen in Übersicht 15.

Übersicht 15: Dokumentationselemente - Fragen und Fragebogen

- Der (Feld-)Fragebogen u. seine stichprobenspezifischen bzw. originalsprachigen Versionen;
- die Position der Frage im Fragebogen (Fragenummer);
- der vollständige Text der Frage;
- der vollständige Text der Antwortkategorien;
- die eindeutige Bedeutung eines erfragten Sachverhalts (z. B. Alter; Einkommen);
- alle Anweisungen für Interviewer bzw. Befragte bei Filterfragen in PAPI Erhebungen bzw. Dokumentation sämtlicher Filterführungen beim Einsatz von CAPI / CATI;
- alle sonstigen Anweisungen für Interviewer oder Befragte;
- alle bei der Befragung eingesetzten Hilfsmittel und Stimuli.

Informationen zur Fragebogenentwicklung sowie die Beschreibung von Pretests und ihren Ergebnisse in entsprechenden Berichten stellen darüber hinaus weiter wichtige Metadaten für Re-Analysen und methodologisch motivierte Sekundäranalysen bereit. Im Rahmen international vergleichender Untersuchungen sind in diesem Zusammenhang auch Informationen über die Übersetzungsprozesse bei der

Erstellung von länderspezifischen Feldfragebögen von besonderer Bedeutung. Ein entsprechendes Vorgehen wird z. B. in den Guidelines and Recommendations zum EVS 2008 im Abschnitt „Translation Process" ausführlich beschrieben (EVS 2008a: 121f).

Gleichzeitig ist die geplante Repräsentation der Daten in strukturierten Datensätzen und ihre präzise Beschreibung eine weitere Voraussetzung, um die Daten auch langfristig zu nutzen.

Übersicht 16: Dokumentationselemente - Variablen und Datenmodifikationen

- Datenfilestruktur und ihrer Variablen;
- Variablennamen und -label;
- Werteausprägungen (Kategorien, Codes) einschließlich aller Arten von fehlenden Angaben;
- Datenmodifikationen in der Datenaufbereitung
- Regeln zur Bildung abgeleiteter Variablen und Gewichtungen.

Diese Metadaten können im Sinne eines Dokumentationsformats mit unterschiedlichen Programmen definiert und erfasst sowie in sehr verschiedenen Dateiformaten gespeichert werden:

- mit einem Editor als textbasierter Codeplan und / oder
- mit einem Statistikprogramm als syntaxbasierte sog. Setupfile und / oder
- unter Verwendung von Textdateien zur Beschreibung aufwendiger Datenmodifikationen oder Bildungsvorschriften einzelner Variablen und / oder
- mit Programmen zur Nutzung des DDI Standards und Speicherung als XML Datei.

Die technische Dokumentation komplexer sozialwissenschaftlicher Daten in Form eines umfassenden Codebuchs und weiteren Materialien zur Studie stellt Nutzern umfassende und hochwertige Informationen bereit, damit das Analysepotential der Daten soweit wie möglich ausgeschöpft werden kann.

Ein Codebuch, Variablenreport oder Datenhandbuch, als breit angelegtes Publikationsformat, integriert folgende Dokumentationsebenen mit den entsprechenden Metadaten:

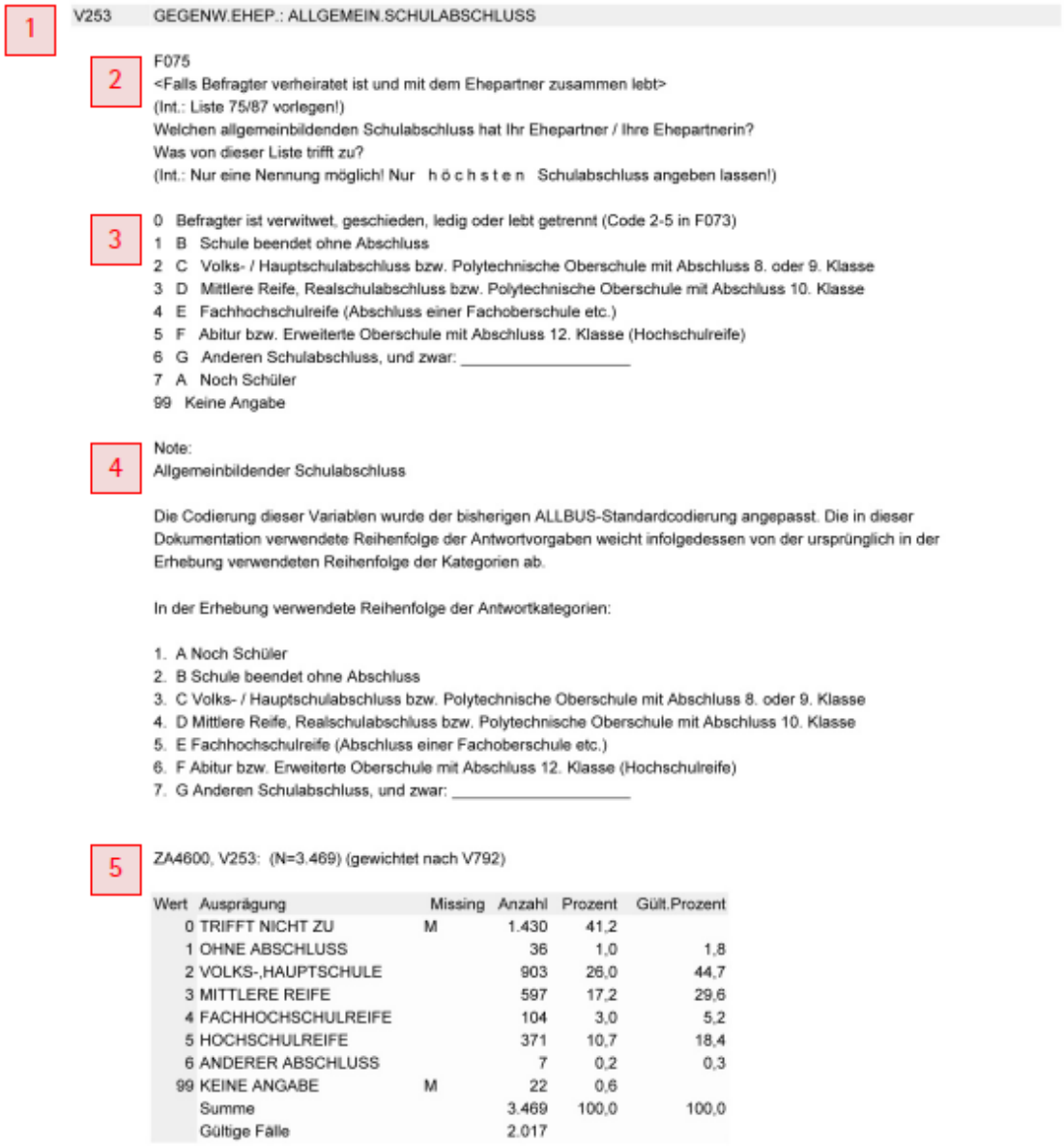
Übersicht 17: Integrierte Dokumentation von und Variablen in einem Codebuch

- Vollständige Frage mit Original-Fragetext und allen weitern Bestandteilen (Anweisungen, Filter, Antwortkategorien etc.) in der Reihenfolge ihrer Verwendung im Fragebogen.
- Bei der Befragung zusätzlich benutzte Stimuli oder Informationen (Listen, Karten).
- Die auf Grundlage der Fragen gebildeten Variablen mit allen Bestandteilen (Namen, Wertetiketten). Wesentlicher Bestandteil ist die vollständige Definition aller fehlenden Werte und eine nachvollziehbare Beschreibung zum Vorgehen bei Recodierungen.
- Alle zusätzlichen gebildeten und abgeleiteten Variablen und Angaben zu ihrer Erstellung.

Umfassendere Anmerkungen (Notes) informieren z. B. über die Bildungs- bzw. Ableitungsvorschriften von konstruierten und harmonisierten Variablen sowie von Gewichtungsvariablen. Gleiches betrifft die Codierung von Angaben zu Beruf oder Bildung, die aufgrund des Umfangs üblicherweise in gesonderten Anhängen dokumentiert werden.

- Anmerkungen über Modifikationen, Abweichungen u. Probleme in Fragen oder Variablen.
- Statistiken, z. B. Häufigkeitsverteilung, Kreuztabellen und / oder statistische Maße.

Abbildung 7: Codebuchseite mit Erläuterungen (Quelle ALLBUS 2008a)



1

V253 GEGENW.EHEP.: ALLGEMEIN.SCHULABSCHLUSS

2

F075
<Falls Befragter verheiratet ist und mit dem Ehepartner zusammen lebt>
(Int.: Liste 75/87 vorlegen!)
Welchen allgemeinbildenden Schulabschluss hat Ihr Ehepartner / Ihre Ehepartnerin?
Was von dieser Liste trifft zu?
(Int.: Nur eine Nennung möglich! Nur h ö c h s t e n Schulabschluss angeben lassen!)

3

0 Befragter ist verwitwet, geschieden, ledig oder lebt getrennt (Code 2-5 in F073)
1 B Schule beendet ohne Abschluss
2 C Volks- / Hauptschulabschluss bzw. Polytechnische Oberschule mit Abschluss 8. oder 9. Klasse
3 D Mittlere Reife, Realschulabschluss bzw. Polytechnische Oberschule mit Abschluss 10. Klasse
4 E Fachhochschulreife (Abschluss einer Fachoberschule etc.)
5 F Abitur bzw. Erweiterte Oberschule mit Abschluss 12. Klasse (Hochschulreife)
6 G Anderen Schulabschluss, und zwar: _____
7 A Noch Schüler
99 Keine Angabe

4

Note:
Allgemeinbildender Schulabschluss

Die Codierung dieser Variablen wurde der bisherigen ALLBUS-Standardcodierung angepasst. Die in dieser Dokumentation verwendete Reihenfolge der Antwortvorgaben weicht infolgedessen von der ursprünglich in der Erhebung verwendeten Reihenfolge der Kategorien ab.

In der Erhebung verwendete Reihenfolge der Antwortkategorien:

1. A Noch Schüler
2. B Schule beendet ohne Abschluss
3. C Volks- / Hauptschulabschluss bzw. Polytechnische Oberschule mit Abschluss 8. oder 9. Klasse
4. D Mittlere Reife, Realschulabschluss bzw. Polytechnische Oberschule mit Abschluss 10. Klasse
5. E Fachhochschulreife (Abschluss einer Fachoberschule etc.)
6. F Abitur bzw. Erweiterte Oberschule mit Abschluss 12. Klasse (Hochschulreife)
7. G Anderen Schulabschluss, und zwar: _____

5

ZA4600, V253: (N=3.469) (gewichtet nach V792)

Wert	Ausprägung	Missing	Anzahl	Prozent	Gült.Prozent
0	TRIFFT NICHT ZU	M	1.430	41,2	
1	OHNE ABSCHLUSS		36	1,0	1,8
2	VOLKS-,HAUPTSCHULE		903	26,0	44,7
3	MITTLERE REIFE		597	17,2	29,6
4	FACHHOCHSCHULREIFE		104	3,0	5,2
5	HOCHSCHULREIFE		371	10,7	18,4
6	ANDERER ABSCHLUSS		7	0,2	0,3
99	KEINE ANGABE	M	22	0,6	
	Summe		3.469	100,0	100,0
	Gültige Fälle		2.017		

Erläuterungen:

- 1: Jeder Frageeinheit der Studie ist eine Variablennummer und ein Variablenlabel eindeutig zugeordnet.
- 2: Bei Variablen, die direkt dem Fragebogen entstammen (Beispiel), steht an dieser Stelle der vollständige Fragetext mit der Fragebogennummer, einschließlich eventueller Interviewer- und Filteranweisungen. Die Notation richtet sich dabei soweit wie möglich nach der Vorlage im Erhebungsinstrument. Bei abgeleiteten oder neu gebildeten Variablen steht an dieser Stelle ein ergänzender Kurzkommentar (Note) zur Variablenbeschreibung.
- 3: Hier stehen die explizit im Datensatz vorhandenen Vercodungen der einzelnen Antwortkategorien sowie die zugehörigen Antworttexte (Volltexte aus den Originalunterlagen). In seltenen Fällen werden Antworttexte ergänzt bzw. Hilfstexte hinzugefügt.

- 4: Weiterführende Informationen stehen direkt nach der Dokumentation der Antwortcodes. Es wird dabei nach Ableitungen der Daten, Bemerkungen und Notes unterschieden:

  Ableitungen der Daten liefern Informationen zu Bildungsvorschriften bei abgeleiteten Variablen. Bemerkungen dienen der Dokumentation von kurzen weiterführenden Informationen. Noten vertiefen das Verständnis der Variablen, indem sie für interessierte Anwender ergänzende Hintergrundinformationen zur Variable liefern.

- 5: Bei den meisten Variablen findet sich an dieser Stelle eine Häufigkeitstabelle. Werteetiketten werden aus dem jeweiligen Datensatz übernommen (vgl. auch die weiteren Anmerkungen zu Gewichtungen im ALLBUS 2008. Variablenreport: xv).

Nicht jedes Forschungsprojekt wird angesichts einer überschaubaren Datenstruktur, des Zeitaufwandes oder knapper Ressourcen ein solch umfassendes Codebuch erstellen wollen oder können.

Es sind aber auch einfachere Codebuchformate denkbar, z. B. indem elektronisch vorliegende Originalfragen und die Variablen in einer Variablenliste in einem Dokument zusammengestellt werden.

Eine solche Dokumentationsform ermöglicht es über den Projektverlauf hinaus, systematisch die Modifikationen der Forschungsdaten und die Bildung neuer Variablen transparent und nachvollziehbar zu dokumentieren.

5.3 Rechtsfragen bei der Archivierung u. Bereitstellung von Studien und Daten

Rechtliche Fragen bei der Archivierung und der Nachnutzung von Forschungsdaten betreffen u. a. die Nutzungsrechte der Daten und den Datenschutz. Spindler und Hillegeist (2011) diskutieren diesbezügliche Fragen im Rahmen der Regelungen des Urheberrechtsgesetzes und der Datenschutzgesetze des Bundes und der Länder. Als Fazit stellen sie fest, „dass der rechtliche Problemschwerpunkt der elektronischen Langzeitarchivierung von Forschungsdaten nicht, wie man zunächst annehmen könnte, urheberrechtlicher, sondern datenschutzrechtlicher Natur ist" (ebd. 69).

Rechtsfragen im Zusammenhang mit dem Urheberrechtsgesetz (UrhG)

Die nachhaltige Sicherung von Forschungsdaten betrifft nicht nur ihre Archivierung, sondern auch die Anforderung, diese Daten Dritten für die wissenschaftliche Forschung und Lehre zur Verfügung zu stellen. Urheberrechtlich relevant ist dabei die Frage, ob die

> „Daten also urheberrechtlich geschützt sind und, sofern dies zutrifft, wer Inhaber der erforderlichen Nutzungsrechte ist." Würde ein solcher Schutz bestehen, dürften die Forschungsdaten nur in ein Archiv überführt werden, „sofern die archivierende Einrichtung Inhaber der erforderlichen Nutzungsrechte wäre bzw. der jeweilige Rechteinhaber der Einrichtung die Archivierung gestatten würde." (ebd. 63)

Dazu stellen die Autoren fest, dass „wissenschaftliche Primärdaten grundsätzlich nicht dem Schutz des Urheberrechtsgesetzes unterliegen" (ebd.), weil es ihnen an der notwendigen geistigen Schöpfungshöhe fehlt, wie es § 2 Abs. 2 UrhG zum Schutz eines Werkes voraussetzt. Der Schutz von Datensammlungen als Datenbankwerke (nach § 4 Abs. 2 UrhG) oder Datenbanken (nach § 87a UrhG) ist im Einzelfall zu prüfen (ebd. 64f).

Aus praktischer Sicht erscheint es bei der Archivierung von Forschungsdaten insgesamt sinnvoll, dass sich eine archivierende Einrichtung zumindest einfache Nutzungsrechte vom Inhaber der Nutzungsrechte einräumen lässt, sofern die Einrichtung diese Rechte nicht selbst besitzt. Um die Daten zu archi-

vieren und öffentlich zugänglich machen zu können, zählen dazu insbesondere das Recht auf Vervielfältigung (UrhG § 16) und das Verbreitungsrecht (§ 17 UrhG) (ebd. 65).

Über die grundsätzlichen Überlegungen der Autoren hinaus ist auch darauf zu achten, bei der Archivierung von studienbezogenen Dokumentationsmaterialien oder auch Sekundärdaten Schutzrechte geistiger Schöpfungen (gemäß Urheberrechtsgesetz) oder gewerblicher Produkte (z. B. Patente, Marken) zu berücksichtigen. Dies gilt beispielsweise für Materialien zu einer Studie, wie etwa Zeitschriftenartikel, Dokumente, Messinstrument(e) u. ä. Materialien, die vervielfältigt und verbreitet werden sollen. Einer entsprechenden Nutzung solcher Materialien dürfen keine Rechte von Dritten entgegenstehen, wie dies z. B. für Kopien aus einem Buch üblicherweise der Fall ist.

Rechtsfragen im Zusammenhang mit dem Datenschutzrecht

Sollen Daten, die für Forschungszwecke etwa mit Methoden der empirischen Sozialforschung erhoben wurden, langfristig archiviert und bereitgestellt werden, sind die datenschutzrechtlichen Anforderungen des Bundes bzw. der Länder Erhebung und Verarbeitung personenbezogener Daten zu beachten. Im Bundesdatenschutzgesetz (BDSG 1990; 2003) werden

- „personenbezogene Daten" nach § 3 Abs. 1 als Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener) definiert. Dazu zählen „auch jene Einzelangaben, die eine bestimmte Person zwar nicht eindeutig oder unmittelbar identifizieren, die es aber erlauben, die Identität der Person mit Hilfe anderer Informationen festzustellen. Man bezeichnet diese als individualisierbare bzw. personenbeziehbare Daten" (Metschke, Wellbrock 2002: 15).

Wie bereits in Kap. 1.3 ausgeführt ist eine wesentliche Voraussetzung für die geplante Datenerhebung und -verarbeitung, dass der Befragte der Nutzung seiner Angaben für wissenschaftliche Zwecke zugestimmt hat (Informierte Einwilligung). Eine zweite Voraussetzung einer geplanten Archivierung und Bereitstellung von Forschungsdaten der empirischen Sozialforschung ist, dass die faktische Anonymisierung sichergestellt ist.

In diesem Zusammenhang sind drei Formen der Anonymisierung zu unterschieden:

- Formales Anonymisieren ist das Entfernen aller direkten Identifikatoren (Namen, Kontaktdaten etc.), z. B. durch Verschlüsselung und Codierung von Befragten (Befragten-ID). Hierbei ist zu unterscheiden, ob die personenbezogene Kontaktdaten für weitere Forschungsvorhaben benötigt werden oder gelöscht werden können.

  Soll der Betroffene auf Grundlage seiner Zustimmung an einer weiteren Befragung etwa im Rahmen eines Panels teilnehmen, sind seine Angaben zu den Merkmalen getrennt von den Kontaktinformationen zu speichern und zu verarbeiten. Dazu werden Betroffene anhand von Pseudonymen verschlüsselt codiert. Das sog. „Pseudonymisieren ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren" (§ 3 Abs. 6a BDSG).

  „Bei der Pseudonymisierung werden die unmittelbar eine Person identifizierenden Daten durch eine für das Einzelvorhaben zu bildende Zuordnungsvorschrift derart verändert, dass das so gebildete Pseudonym nur mit Kenntnis dieser Zuordnungsvorschrift wieder einer natürlichen Person zugeordnet werden kann." (Metschke, Wellbrock 2002: 23).

  Werden z. B. Kontaktdaten nicht mehr für Zwecke des Forschungsvorhabens benötigt, sollen sie so früh wie möglich gelöscht werden, um einer faktischen Anonymisierung näher zu kommen.

Der Umfang an erhobenen Merkmalen sowie regionalen oder sonstigen personalisierbare Zuordnungen bleibt jedoch bei der formalen Anonymisierung zunächst erhalten, soweit keine weiteren Anonymisierungsmethoden angewendet werden.

- Erst der Einsatz weiterer, oftmals kombinierter Verfahren, führt zu einer **faktischen Anonymisierung**.

  Dazu gehören u. a. die Kategorisierung von Berufsangaben (ISCO Codierung), die Vergröberung von Anwortkategorien (z. B. Bildung von Altersgruppen) und die Verallgemeinerung von Orts- bzw. Regionalangaben.

  So ist faktisches Anonymisieren „das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können" (§ 3 Abs. 6 BDSG). Faktisch anonymisierte Einzeldaten werden von Datenserviceeinrichtungen auch als „Scientific Use Files" gekennzeichnet.

- Eine **absolute Anonymisierung** liegt vor, wenn die Daten in einer Weise angepasst wurden, so dass eine direkte oder indirekte Identifizierung einer Person auf jeden Fall ausgeschlossen werden kann. Absolut anonymisierte Einzeldaten werden oftmals auch als „Public Use Files" bezeichnet.

Im Zuge der Archivierung von quantitativen Umfragedaten einer einzelnen Erhebung (Querschnittsstudie) liegen üblicherweise keine Informationen mehr vor, die eine Person unmittelbar identifizierbar macht (vergliche dazu auch Quandt, Mauer 2012:73). Trotzdem besteht, je nach Datenlage, die Notwendigkeit, weitere Maßnahmen zur faktischen Anonymisierung (Metschke, Wellbrock 2002: 21f) zu prüfen und soweit notwendig durchzuführen, um dem Restrisiko einer mittelbaren Re-Identifizierung auf Grundlage (kombinierbarer) persönlicher demographischer Angaben entgegen zu wirken. Ein entsprechendes Vorgehen sollte deshalb frühzeitig zwischen Forschungsprojekt und der Einrichtung abgestimmt werden, die die Daten am Projektende aufnehmen und bereitstellen soll.

Eine systematische Beschreibung zum „Datenschutz und die Archivierung von Daten der qualitativen empirischen Sozialforschung" leistet der Beitrag von Watteler (2010b). Anhand von drei Phasen (Datenerhebung, Archivierung, Sekundärnutzung) wird beschrieben, welche Maßnahmen zum Schutz von personenbezogenen Daten entlang der wissenschaftlichen Wertschöpfungskette ergriffen werden müssen. Dazu werden kontextbezogen die einschlägigen Gesetze aufgeführt, auf deren Grundlage spezielle datenschutzrechtliche Vorgaben zu berücksichtigen sind.

Beim Umgang mit personenbeziehbaren Daten muss eine archivierende Einrichtung darüber hinaus die datenschutzrechtlichen Vorschriften hinsichtlich der technischen und organisatorischen Schutzmaßnahmen bei der Datenverarbeitung berücksichtigen. Diese betreffen insbesondere die Regelung des kontrollierten und autorisierten Zugangs zu den technischen Systemen sowie des Zugriffs und die Bearbeitung dieser Daten (vgl. § 9 BDSG mit Anlage zu § 9 Satz 1; Datenschutzgesetze der Länder).

Die Weitergabe anonymisierter Daten an Dritte unterliegt nicht mehr den Regelungen des Datenschutzes und ist dann auch forschungsethisch unkritisch. Vor diesem Hintergrund sind datenschutzrechtliche Aspekte auch bei der Gestaltung eines Nutzungs- bzw. Archivierungsvertrages zu berücksichtigen.

5.4 Rechtemanagement: Der Archivierungsvertrag des GESIS Datenarchivs

Neben der Frage, wo und wie die Daten und deren Dokumentation gesichert werden, sind im Forschungsprojekt auch Aspekte des zukünftigen Rechtemanagements hinsichtlich der produzierten Forschungsdaten zu bedenken. Diese Thematik ist erfahrungsgemäß sowohl für Forscher im Projekt als auch für die zukünftige datenhaltende Einrichtung von Relevanz. D. h., es ist etwa zu klären, zu welchem Zeitpunkt, zu welchem Zweck und unter welchen Bedingungen der Zugang zur Studie und den Forschungsdaten ermöglicht werden soll.

Je nachdem, an welche datenhaltende Einrichtung die Forschungsdaten zur Sicherung und Bereitstellung übergeben werden, ist eine vertragliche Regelung der Nutzungs- und Archivierungsrechte empfehlenswert.. Dabei können etwa Zugangsregelungen, Aufbewahrungsfristen und Nutzungsbedingungen vereinbart werden. Darüber hinaus wäre zu überlegen, ob ein Ansprechpartner für Fragen bezüglich der Studie und der Daten auch nach Projektende zur Verfügung steht.

Bei einer geplanten Langfristsicherung im GESIS Datenarchiv werden Nutzungsrechte auf der Basis von Zugangskategorien vertraglich geregelt, wie sie auch in internationalen Archivkontexten angewendet werden. Dies sind die Kategorien zur Nutzung der Daten

- ohne Einschränkung, oder
- nur für wissenschaftliche Zwecke, oder
- nur für wissenschaftliche Zwecke mit Information des Datengebers bei Publikationen, oder
- nur für wissenschaftliche Zwecke mit schriftlicher Weitergabegenehmigung durch den Datengeber (vgl. GESIS Datenarchiv 2007: Benutzungsordnung).

Die langfristige Archivierung und Bereitstellung von Forschungsdaten wird durch einen Archivierungsvertrag zwischen dem GESIS Datenarchiv und dem Datengeber geregelt. Die zentralen Aspekte, die durch einen solchen Vertrag vereinbart werden, zeigt die folgende Übersicht.

Übersicht 18: Zentrale Aspekte eines Archivierungsvertrages

enstand der Vereinbarung ist die Bereitstellung von Studie(n) durch den Datengeber mit Daten und en, die zur Interpretation der Daten notwendig erscheinen.

2. Verfügungs- und Nutzungsrechte

Der Datengeber überträgt dem Archiv das Recht,

- Daten und Texte systematisch zu archivieren,
- zum Zweck der physischen Sicherung und weiterer Auswertungen alle zweckdienlichen technischen Mittel, Formate und Methoden einzusetzen,.
- Texte zu digitalisieren und Daten und Texte gemäß der festgelegten (Nutzungs- bzw. Weitergabe-)Kategorie bereitzustellen.
- Für die genannten Zwecke überträgt der Datengeber dem Archiv alle notwendigen Nutzungsrechte, im Besonderen das Vervielfältigungsrecht (§ 16 UrhG) sowie das Recht der öffentlichen Zugänglichmachung (§ 19 a UrhG). Die Einräumung des Nutzungsrechts erfolgt räumlich und zeitlich unbeschränkt. Der Datengeber willigt in die Veröffentlichung der Metadaten der Studien ein.

Das Archiv übernimmt die kostenlose Aufbewahrung der gelieferten Materialien und stellt sie dem Datengeber auf einfache Anfrage zur Verfügung.

3. Gewährleistungen

Der Datengeber gewährleistet, dass er zur Einräumung von Nutzungsrechten berechtigt ist und dass einer vertragsgemäßen Nutzung der Materialien keine Rechten Dritter entgegenstehen.

Der Datengeber ist damit einverstanden, dass die Materialien für Sekundäranalysen durch eigenständige wissenschaftliche Untersuchungen und unabhängig von den eigenen Forschungszielen ausgewertet werden können.

4. Archivierung: Zur Bearbeitung und Aufnahme in den Datenbestandskatalog wird die Bezeichnung der Studie durch die Namen der Primärforscher und den Studientitel schriftlich festgelegt.
5. Haftung: Das Archiv haftet nur bei Vorsatz und grober Fahrlässigkeit bei Archivierungstätigkeiten zur Ausführung der vertraglichen Vereinbarungen.
6. Rechtsnachfolge: Sämtliche Rechte an der archivierten Studie gehen in folgenden Fällen treuhänderische an das Archiv über:
   - Ableben des Datengebers,
   - Schließung der datengebenden Institution,
   - Nicht-Nachvollziehbarkeit des Verbleibs des Datengebers,
   - oder wenn eine Rechtsnachfolge nicht mehr einwandfrei nachvollziehbar ist.
7. Datenschutz

Archiv und Datengeber verpflichten sich, im Zusammenhang mit der Durchführung dieser Vereinbarung die einschlägigen Datenschutzbestimmungen einzuhalten. Der Datengeber erklärt insbesondere, etwaige Datenschutzbestimmungen anderer Länder, in denen die Daten erhoben wurden, im Zusammenhang mit der Nutzungseinräumung beachtet zu haben. Das Archiv behält sich bei datenschutzrechtlichen Bedenken vor, geeignete Anonymisierungsmaßnahmen vorzuschlagen.

8. Laufzeit: Die Vereinbarung wird auf unbestimmte Zeit geschlossen, soweit nichts anderes ausdrücklich vereinbart wurde.

5.5 Auswahl und Übergabe der zu sichernden Daten und Dokumentationen

Das Forschungsvorhaben entscheidet am Projektende, welche Daten, Metadatendokumentationen und andere Projektergebnisse zur langfristigen Sicherung und Bereitstellung übergeben werden sollen.

Hierbei sollten neben den Daten alle erforderlichen Materialien berücksichtigt werden, die den Entstehungszusammenhang und die Bedeutung der Daten verständlich und nachvollziehbar machen. Dazu zählen neben der

- Endversion der Daten, möglichst aufbereitet für die Verwendung in einem Statistikprogramm,

mindestens auch

- die Dokumentation der Daten (Codeplan bei ASCII Daten; Codebuch oder Ähnliches),
- der Originalfeldfragebogen bzw. ein entsprechendes dokumentiertes Messinstrument einschließlich aller zusätzlich eingesetzten Hilfsmittel (Show-Card, Interviewerinstruktionen o. ä.) und
- die Beschreibung von Methodendesign und Datenerhebung (Methodenbericht; Feldbericht).

Weiterhin ist es empfehlenswert, zusätzliche Materialien, Berichte und Dokumentationen des Forschungsprojektes zu archivieren, soweit sie für eine Sekundäranalyse hilfreich sein können. Nach der Auswahl aller zu sichernden Daten- und Dokumentationsdateien sollten diese inhaltlich und formal kontrolliert und fehlende Informationen ergänzt werden. Dabei sollten auch alle rechtlichen Aspekte (Datenschutz, Nutzungsrechte) abschließend geprüft werden.

Die Datendateien sowie Projekt-, Methoden- und Datendokumentation sind dann in den abgesprochenen Dateiformaten zur Übergabe vorzubereiten. Zur allgemeinen Speicherung und Weitergabe der Daten zur Archivierung können die Dateiformate der üblicherweise benutzten Statistiksoftware (SAS, SPSS, STATA) bzw. ASCII oder Tab-getrennte Dateiformate (TXT, CSV) benutzt werden. Gleiches gilt für die Dateiformate zur Erstellung von Dokumentation mit gängigen Softwarepaketen. Gegebenenfalls werden diese Daten dann in der datensichernden Einrichtung zusätzlich in gängige Formate für die längerfristige Sicherung der Studienmaterialien konvertiert (z. B. ASCII, PDF, XML).

Die zur Archivierung vorgesehenen Materialien werden dann zum vereinbarten Zeitpunkt an die datensichernde Einrichtung übergeben. Sinnvollerweise wird der Eingang aller Materialien der Studie durch ein Eingangsprotokoll oder Ähnliches dokumentiert und bestätigt. Die Aufnahme, Kontrolle, Archivierung und Bereitstellung der Studie erfolgt dann im Weiteren auf Grundlage der Arbeitsprozesse und Standards der archivierenden Serviceeinrichtung (vgl. dazu auch Quandt, Mauer 2012: 70).

6 Dienstleistungen des GESIS Datenarchivs zur Langzeitarchivierung von sozialwissenschaftlichen Forschungsdaten

Das GESIS Datenarchiv für Sozialwissenschaften ist ein Digitales Langzeitarchiv und Dienstleister für Primärforscher zur Sicherung, Dokumentation, Aufwertung und Bereitstellung ihrer Daten (vgl. Mauer 2011; Quandt, Mauer 2012: 68ff). Zu den Aufgaben zählen die:

- Bereitstellung und Unterhaltung notwendiger Dienste, um Sekundäranalysen, Replikationen sowie vergleichende Untersuchungen über Raum und Zeit zu ermöglichen und den Zugang zu internationalen Forschungsdaten zu gewährleisten;
- Langzeitarchivierung von Forschungsdaten, um die Langzeitverfügbarkeit und Interpretierbarkeit der Daten zu sichern.

Die Übernahme der Verantwortung für die Langzeitarchivierung von sozialwissenschaftlichen Forschungsdaten sowie die Nutzungs- und Zielgruppenorientierung ist in der Satzung der GESIS – Leibniz-Institut für Sozialwissenschaften (2010) verankert.

Das Datenarchiv für Sozialwissenschaften erfüllt seine Aufgaben durch verschiedene wissenschaftliche und technische Dienstleistungen, die im Folgenden vorgestellt werden.

Information und Beratung von Primärforschern und Datenprojekten

Ziel der Akquisitionsaktivitäten und Beratungsangebote des GESIS Datenarchivs ist es u. a. möglichst früh im Lebenszyklus von Forschungsdaten anzusetzen und Datengeber bei der Klärung bzgl. Umfang, Formaten, Aufbereitungs- und Dokumentationszielen sowie bei Nutzungsrechten und Zugangsklassen zu unterstützen. Dazu bietet das Datenarchiv Beratungen, Schulungen und Informationen zum Thema Datenmanagement und Datenarchivierung an. Die Archivierung von Studien wird durch den Abschluss eines Archivierungsvertrages mit dem Datengeber rechtlich transparent und verbindlich geregelt.

da|ra – Registrierungsagentur für sozialwissenschaftliche Forschungsdaten

Forschungsdaten und Datenproduzenten können ihre sozialwissenschaftlichen Forschungsdaten in der Registrierungsagentur da|ra erfassen lassen, um die eigenen Datenquellen dauerhaft zitierfähig und leicht zugänglich zu gestalten.

Aufnahme von Studien ins das Archiv

Neu aufgenommene Studien durchlaufen verschiedene Verfahren und Dienste der Standardarchivierung. Dazu zählen:

- die Prüfung der Vollständigkeit, Übereinstimmung und Nutzbarkeit des übergebenen Materials (Daten, Messinstrumente und Dokumente),
- die technische Kontrolle der Dateien (Formate, Lesbarkeit, Virenfreiheit etc.) und
- die Konsistenzprüfungen der Daten und Prüfungen zum Datenschutz und ggf. korrigierende Aufbereitungen in Abstimmung mit dem Datengeber.

Die Standardverfahren bezüglich aller Studien beinhalten neben den Eingangskontrollen folgende Maßnahmen:

- Formale und inhaltliche Erschließung der neu aufgenommenen Studie durch die Studienbeschreibung. Neben der Vergabe einer zentralen Studienummer als Identifier des Datenbestan-

des werden die inhaltlichen, methodischen und technischen Charakteristika der Studie beschrieben.

- Versionierung der Daten mit Errata-Historie und Vergabe eines dauerhaften Identifikators je Datensatzversion (DOI – Digital Object Identifier) und Registrierung des Datensatzes in da|ra.
- Erstellen eines Archiverungspaketes, das alle Originale, aufbereite Versionen von Daten und Dokumenten, normalisierte Dateien, sowie die dazugehörigen Metadaten beinhaltet.
- Erzeugung der für den Service bestimmten Daten und Dokumente in nutzerfreundlichen Datei- und Publikationsformaten.

Mehrwertdienste in der Datenaufbereitung und Dokumentation

Für ausgewählte Studien oder Studienkollektionen erfolgt in Abstimmung mit den Trägern von Umfrageprogrammen oder einzelnen Forschungsprojekten eine mehrwertorientierte Aufbereitung und Dokumentation komplexer Daten. Die Mehrwertdienste umfassen die

- Erstellung von umfassenden Metadaten auf Studien-, Fragen- und Variablenebene,
- Standardisierung und Harmonisierung von Variablen,
- Ergänzungen von Kontextinformationen und Aggregatdaten,
- Integration bzw. Kumulation vergleichender Daten und Dokumentation der Datenaufbereitung und die
- Veröffentlichung in verschiedenen Publikationsformaten (Variablenreport, Methodenbericht, Technical Report).

Datenbereitstellung und Nutzerberatung

Der Datenservice schließt die verschiedenen Facetten von der zentralen Nutzerberatung bis hin zu speziellen Dienstleistungen der GESIS Forschungsdatenzentren ein. Das GESIS Datenarchiv vermittelt auch den Zugang zu Datenbeständen ausländischer Archive über die Archivnetzwerke ICPSR, CESSDA und IFDO (siehe Anhang A.1).

Zu den technischen Diensten zur Erschließung und Bereitstellung von Daten und Dokumentationen zählen u. a. die

- Veröffentlichung von Studien auf Basis vereinbarter Nutzungsrechte und des Datenschutzes;
- Erschließung aller Studien durch Recherchen in umfassenden Metadatenbeständen;
- Bereitstellung von studienbezogenen Materialien mit Vorschlägen zur bibliographischen Zitation des Datenfiles sowie Angaben zu datenbezogener Primär- und Sekundärliteratur.

Der Zugang zu den Daten und Metadaten wird über die Online-Portale DBK (Datenbestandskatalog), ZACAT (Online Study Catalogue) und HISTAT (Datenbank Historische Statistik) angeboten. Individuelle Bestellungen von Daten und Studienmaterial werden auf Datenträgern oder per ftp bereitgestellt.

Langfristsicherung und Langzeitarchivierung

Die Archivierung einer Studie erfolgt nach erfolgreichem Abschluss aller Eingangskontrollen durch die Überführung aller zur Studie gehörenden Objekte in das zentrale Archivsystem. Die langfristige Substanzerhaltung wird u. a. durch eine räumlich getrennte, redundante Datenhaltung, den Einsatz unterschiedlicher Speichertechniken sowie langfristig angelegte Migrationsstrategien gesichert.

A. Anhang

Zugriff auf alle Links im Anhang A.1 bis A.4 zuletzt am 16.09.2012

A.1 Nationale und internationale Datenquellen für Sekundäranalysen

GESIS – Datenarchiv für Sozialwissenschaften

- GESIS Angebote im Forschungsdatenzyklus:
  http://www.gesis.org/unser-angebot/
- DBK - Datenbestandskatalog: Recherche in den Studienschreibungen der im Archiv gespeicherten Studien und Primärdaten und Download von Materialien (Variablenreports, Berichte, Fragebogen o. ä.):
  http://www.gesis.org/unser-angebot/recherchieren/datenbestandskatalog/
- ZACAT - Online Study Catalogue: Recherche (bis auf Variablenebene), Datenanalyse und Download von Umfragedaten aus speziellen Studienkollektionen:
  http://www.gesis.org/unser-angebot/recherchieren/zacat-online-study-catalogue/
- HISTAT - Zeitreihen der Historischen Statistik: Recherche und Download:
  http://www.histat.gesis.org/index.php
- Angebot zur Datenarchivierung:
  http://www.gesis.org/unser-angebot/archivieren-und-registrieren/datenarchivierung/
- Registrierungsagentur da|ra für Sozial- und Wirtschaftsdaten:
  http://www.da-ra.de/
- Sammlung nationaler und internationaler Datenquellen und Kontextinformationen:
  - http://www.gesis.org/das-institut/kompetenzzentren/european-data-laboratory/data-resources/context-information/
  - http://www.gesis.org/eurobarometer/service-guide/weblinks/

RatSWD, Rat für Sozial- und Wirtschaftsdaten

Deutsche Forschungsdaten- und Datenservicezentren:
http://www.ratswd.de/dat/fdz.php

CESSDA, Council of European Social Science Data Archives

http://www.cessda.org/about/members/index.html

ICPSR, Inter-University Consortium for Political and Social Research

http://www.icpsr.umich.edu

IFDO, International Federation of Data Organizations

http://www.ifdo.org/

Sociosite, Social Science Information System (University of Amsterdam)

http://www.sociosite.net/index.php

A.2 Literatur und Referenzen zum Datenmanagement

AAPOR, 2008. The American Association for Public Opinion Research. 2008. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 5th edition. Lenexa, Kansas: AAPOR: http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions&Template=/CM/ContentDisplay.cfm&ContentID=1273

ADM, Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V.:

- Homepage: http://www.adm-ev.de/index.php?id=startseite
- ICC/ESOMAR Kodex: http://www.adm-ev.de/index.php?id=kodex
- Richtlinien: http://www.adm-ev.de/index.php?id=richtlinien

  Vergleiche auch ISO 20252:2006 und ISO 26362:2009in Anhang A.3

ADM, Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e. V., o. J. Erklärung zum Datenschutz und zur absoluten Vertraulichkeit Ihrer Angaben bei mündlichen oder schriftlichen Interviews: http://www.adm-ev.de/fileadmin/user_upload/PDFS/Merkblatt.pdf

Akremi, Leila; Baur, Nina; Fromm, Sabine (Hrsg.), 2011. Datenanalyse mit SPSS für Fortgeschrittene 1. Datenaufbereitung und uni- und bivariate Statistik. 3. Aufl., Wiesbaden.

ALLBUS, Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, 2008a. Variable Report (ZA4600). Terwey, Michael; Baltzer, Stefan. GESIS Variable Reports Nr. 2011/04: http://info1.gesis.org/dbksearch/file.asp?file=ZA4600_cdb.pdf

ALLBUS, Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, 2008b. Studienbeschreibung (ZA4600): http://dx.doi.org/doi:10.4232/1.10834

Allianz 2010: Allianz der deutschen Wissenschaftsorganisationen. Grundsätze zum Umgang mit Forschungsdaten: http://www.allianzinitiative.de/fileadmin/user_upload/Home/Video/Grunds%C3%A4tze%20Umgang%20mit%20Forschungsdaten.pdf

ASI, Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V.: www.asi-ev.org/

Bauske, Franz 2000. Das Studienbeschreibungsschema des Zentralarchivs. In ZA-Information Nr.47, S. 73-80: http://www.gesis.org/fileadmin/upload/forschung/publikationen/zeitschriften/za_information/ZA-Info-47.pdf

BDSG, Bundesdatenschutzgesetz, 1990. Stand: Neugefasst durch Bek. v. 14.1.2003 I 66; zuletzt geändert durch Art. 1 G v. 14.8.2009 I 2814: http://www.gesetze-im-internet.de/bdsg_1990/

Brislinger, Evelyn et al., 2009. Empfehlungen - Datenaufbereitung und Dokumentation. Arbeitspapier GESIS Datenarchiv für Sozialwissenschaften (unveröffentlicht).

Brislinger, Evelyn et al., 2011. European Values Study 2008. Project and Data Management. GESIS Technical Reports 2011/14: http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2011/TechnicalReport_2011-14.pdf

Bundesstatistikgesetz – BstatG.
Bundesstatistikgesetz vom 22. Januar 1987 (BGBl. I S. 462, 565), das zuletzt durch Artikel 3 des Gesetzes vom 7. September 2007 (BGBl. I S. 2246) geändert worden ist
http://www.gesetze-im-internet.de/bundesrecht/bstatg_1987/gesamt.pdf

BVerfGE 65, 1. Urteil des Ersten Senats des Bundesverfassungsgerichts vom 15. Dezember 1983 - 1 BvR 209/83 u. a. - sog. Volkszählungsurteil. Die Leitsätze der Entscheidung:
http://sorminiserv.unibe.ch:8080/tools/ainfo.exe?Command=ShowPrintText&Name=bv065001

DCC, Digital Curation Center, 2010. Data Management Plans:
http://www.dcc.ac.uk/resources/data-management-plans

Destatis, Statistischen Bundesamt:
https://www.destatis.de/

DFG, Deutsche Forschungsgemeinschaft, 1998. Denkschrift: Sicherung guter wissenschaftlicher Praxis. Weinheim:
http://www.dfg.de/download/pdf/dfg_im_profil/reden_stellungnahmen/download/empfehlung_wiss_praxis_0198.pdf

DFG, Deutsche Forschungsgemeinschaft, 2009. Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten:
http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf

DFG, Deutsche Forschungsgemeinschaft, 2011. 1.01 Ergänzender Leitfaden für die Beantragung von Langfristvorhaben [10/11]:
http://www.dfg.de/formulare/1_01/1_01.pdf

DFG, Deutsche Forschungsgemeinschaft, 2012a. Informationsmanagement und Informationsinfrastruktur in Sonderforschungsbereichen
http://www.dfg.de/foerderung/programme/koordinierte_programme/sfb/module/modul_inf/

DFG, Deutsche Forschungsgemeinschaft, 2012b. 50.06 Merkblatt Sonderforschungsbereiche [06/12]:
http://www.dfg.de/formulare/50_06/50_06_de.pdf

DFG, Deutsche Forschungsgemeinschaft, 2012c. 54.01 Leitfaden für die Antragstellung - Projektanträge [06/12]:
http://www.dfg.de/formulare/54_01/54_01_de.pdf

EU, European Union, 2010. Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High level Expert Group on Scientific Data. A submission to the European Commission. 2010:
http://ec.europa.eu/information_society/newsroom/cf/document.cfm?action=display&doc_id=707

Eurobarometer 50.0. European Parliament and Radioactive Waste. October-November 1998 (ZA3085 / ICPSR 2330). Codebook 1st. ZA Edition, 2003:
http://info1.gesis.org/dbksearch/file.asp?file=ZA3085_cdb.pdf

EVS, European Values Study, 2011. EVS 1981-2008 Variable Report Longitudinal Data File. Documentation release 2011/12/30. GESIS-Variable Reports No. 2011/10:
https://info1.gesis.org/dbksearch18/download.asp?id=19015

EVS, European Values Study, 2008a. Guidelines and Recommendations. Documentation of the full data release (Integrated Dataset, Archive-Study-No. ZA4800). GESIS Technical Reports No. 2010/16:
http://info1.gesis.org/dbksearch/file.asp?file=ZA4800_standards.pdf

EVS, European Values Study, 2008b. Variable Report. Integrated Dataset. Archive Study No. 4800 (v3.0.0):
http://dx.doi.org/doi:10.4232/1.11004

EVS, European Values Study, 1999-2000. Extended variable overview EVS1981-2008: V154 & V154_5C:
http://info1.gesis.org/evs/variables/compview.asp?db=QEVSLF&id=ZA3811&all=&lang=en&id2=&var2=&lang2=en&tvsearch=V154_5C&ts1=1&ts2=1&ts3=1&var=V154_5C

GESIS Datenarchiv für Sozialwissenschaften, 2007. Benutzungsordnung.
http://www.gesis.org/unser-angebot/daten-analysieren/datenservice/benutzungsordnung/

GESIS – Leibniz-Institut für Sozialwissenschaften e. V., 2010. GESIS Satzung,:
http://www.gesis.org/das-institut/der-verein/satzung/

Häder, Michael, 2010. Empirische Sozialforschung. 2. Aufl., Wiesbaden.

Häder, Michael, 2009. Der Datenschutz in den Sozialwissenschaften. Anmerkungen zur Praxis sozialwissenschaftlicher Erhebungen und Datenverarbeitung in Deutschland. RatSWD Working Paper No. 90:
http://www.ratswd.de/download/RatSWD_WP_2009/RatSWD_WP_90.pdf

ICPSR, Inter-University Consortium for Political and Social Research, 2012. Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle. 5th ed. Ann Arbor, MI.:
http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

ISSP, International Social Survey Programme, 2010a. Standard Setup for ISSP 2010 (Environment III):
http://www.gesis.org/fileadmin/upload/dienstleistung/daten/umfragedaten/issp/members/setups/StandardSetupEnvironment_2010_20110127.sps

ISSP, International Social Survey Programme, 2010b. Background Variables & Further Coding Standards:
http://www.gesis.org/issp/issp-members-area/coding-standards/

Jensen, Uwe, 2011. Datenmanagementpläne. In: Büttner, Stephan; Hobohm, Hans-Christoph; Müller, Lars (Hrsg.). Handbuch Forschungsdatenmanagement. 2011, Bad Honnef, S. 71-82:
http://opus4.kobv.de/opus4-fhpotsdam/files/208/HandbuchForschungsdatenmanagement.pdf

Jensen, Uwe; Katsanidou, Alexia; Zenk-Möltgen, Wolfgang, 2011. Metadaten und Standards. In: Büttner, Stephan; Hobohm, Hans-Christoph; Müller, Lars (Hrsg.). Handbuch Forschungsdatenmanagement. 2011, Bad Honnef, S. 83-100:
http://opus4.kobv.de/opus4-fhpotsdam/files/208/HandbuchForschungsdatenmanagement.pdf

Kromrey, Helmut, 2009. Empirische Sozialforschung. 12. Aufl., Stuttgart.

Lück, Detlev; Baur, Nina, 2011. Vom Fragebogen zum Datensatz. In: Akremi, Leila; Baur, Nina; Fromm, Sabine (Hrsg.). Datenanalyse mit SPSS für Fortgeschrittene 1. Datenaufbereitung und uni- und bivariate Statistik. 3. Aufl., Wiesbaden, S. 22-58.

Lück, Detlev, 2011. Mängel im Datensatz beseitigen. In: Akremi, Leila; Baur, Nina; Fromm, Sabine (Hrsg.). Datenanalyse mit SPSS für Fortgeschrittene 1. Datenaufbereitung und uni- und bivariate Statistik. 2011; 3. Aufl., Wiesbaden, S. 66-80.

Mauer, Reiner, 2011. GESIS Datenarchiv für Sozialwissenschaften. Vortrag anlässlich des Workshop „Archivierung sozial-und wirtschaftswissenschaftlicher Datenbestände". Deutsche Nationalbibliothek, Frankfurt, 15./16. September 2011:
http://www.ratswd.de/ver/docs_Archivierung_2011/mauer.pdf

Metschke, Rainer; Wellbrock Rita, 2002. Datenschutz in Wissenschaft und Forschung. Materialien zum Datenschutz Nr. 28., 3. Aufl. Berlin, 2002:
http://www.datenschutz-berlin.de/attachments/47/Materialien28.pdf?1166527077

NSF, National Science Foundation, 2010. Social, Behavioral and Economic Sciences Directorate (SBE). Data management plan:
http://www.nsf.gov/sbe/SBE_DataMgmtPlanPolicy.pdf

NSF, National Science Foundation, 2011. Dissemination and Sharing of Research Results:
http://www.nsf.gov/bfa/dias/policy/dmp.jsp

OECD, Organisation for Economic Co-Operation and Development, 2007. OECD Principles and Guidelines for Access to Research Data from Public Funding:
http://www.oecd.org/dataoecd/9/61/38500813.pdf

Quandt, Markus; Mauer, Reiner, 2012. Sozialwissenschaften. In: Neuroth, Heike; Strathmann, Stefan Oßwald, Achim; Scheffel, Regine; Klump, Jens; Ludwig, Jens (Hrsg.), Langzeitarchivierung von Forschungsdaten: Eine Bestandsaufnahme. 2012, Göttingen, S. 61-81:
http://nestor.sub.uni-goettingen.de/bestandsaufnahme/nestor_lza_forschungsdaten_bestandsaufnahme.pdf

Schnell, Rainer; Hill, Paul B.; Esser, Elke, 2011. Methoden der empirischen Sozialforschung. 9. Aufl., München.

Schroeder, Kathrin, 2010: 9.4 Persistent Identifier (PI) – ein Überblick. In: H. Neuroth et al., Hrsg. 2010. NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. (Version 2.3):
http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949

Spindler, Gerald; Hillegeist, Tobias, 2011. Rechtliche Probleme der elektronischen Langzeitarchivierung von Forschungsdaten. In: Büttner, Stephan; Hobohm, Hans-Christoph; Müller, Lars (Hrsg.). Handbuch Forschungsdatenmanagement. 2011, Bad Honnef, S. 63-69.
http://opus4.kobv.de/opus4-fhpotsdam/files/208/HandbuchForschungsdatenmanagement.pdf

UKDA, UK Data Archive, 2011. Van den Eynden, Verle et al., 2011. Managing and Sharing Data: Best Practice For Researchers. 3. ed.:
http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

UrhG, Urheberrechtsgesetz. Gesetz über Urheberrecht und verwandte Schutzrechte:
http://www.gesetze-im-internet.de/urhg/index.html

Watteler, Oliver, 2010a. Erstellung von Methodenberichten für die Archivierung von Forschungsdaten. Arbeitspapier GESIS Datenarchiv für Sozialwissenschaften (unveröffentlicht).

Watteler, Oliver, 2010b: Datenschutz und die Archivierung von Daten in der qualitativen empirischen Sozialforschung. In: Medjedović, Irena; Witzel, Andreas (Hrsg.): Wiederverwendung qualitativer Daten: Archivierung und Sekundärnutzung qualitativer Interviewtranskripte. 2010, Wiesbaden, VS Verl. für Sozialwiss., S. 55-94.

Wittenberg, Reinhard; Cramer, Hans, 2003. Datenanalyse mit SPSS für Windows, 3. Aufl., Stuttgart.

Zehntes Sozialgesetzbuch - Sozialverwaltungsverfahren und Sozialdatenschutz - (SGB X)
"Zehntes Buch Sozialgesetzbuch - Sozialverwaltungsverfahren und Sozialdatenschutz - (Artikel 1 des Gesetzes vom 18. August 1980, BGBl. I S. 1469 und Artikel 1 des Gesetzes vom 4. November 1982, BGBl. I S. 1450) in der Fassung der Bekanntmachung vom 18. Januar 2001 (BGBl. I S. 130), das zuletzt durch Artikel 8 des Gesetzes vom 21. Juli 2012 (BGBl. I S. 1566) geändert worden ist
http://www.gesetze-im-internet.de/bundesrecht/sgb_10/gesamt.pdf

Zenk-Möltgen, Wolfgang, Habbel, Norma, 2012. GESIS-Technical Reports 2012|01. Der GESIS Datenbestandskatalogs und sein Metadatenschema. Version 1.8.
http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_methodenberichte/2012/TechnicalReport_2012-01.pdf

A.3 Sozialwissenschaftlich relevante Standards und Klassifikationen

Qualitätskriterien in der der Umfrageforschung

ISO 20252:2006. Markt-, Meinungs- und Sozialforschung. Vokabular und Anforderungen an die Dienstleistung: http://www.iso.org/iso/catalogue_detail?csnumber=39339

- Vergleiche auch: http://www.umfragen.info/online/umfrage/2006/iso-202522006-markt-meinungs-und-sozialforschung/
- Vergleiche auch Richtlinien und Kodex des ADM in Anhang A.2

ISO 26362:2009: Access panels in market, opinion and social research Vocabulary and service requirements:
http://www.iso.org/iso/catalogue_detail?csnumber=43521

Sozialwissenschaftliche Skalen und Standarddemographie

Demographischen Standards, 2010. Gemeinsame Empfehlung von ADM, ASI und Statistischem Bundesamt:
https://www.destatis.de/DE/Methoden/DemografischeRegionaleStandards/Standards.html

- Fragebogenversionen:
https://www.destatis.de/DE/Methoden/DemografischeRegionaleStandards/DemografischeStandardsInfo.html?nn=173768

EHES, 2010. Elektronisches Handbuch zu Erhebungsinstrumenten im Suchtbereich, Version 4.00, 2010:
http://www.gesis.org/unser-angebot/studien-planen/zis-ehes/ehes/

- Download:
http://www.gesis.org/unser-angebot/studien-planen/zis-ehes/

ZIS, 2010. Zusammenstellung sozialwissenschaftlicher Items und Skalen, ZIS Version 14.00, 2010:
http://www.gesis.org/unser-angebot/studien-planen/zis-ehes/zis/

- Download:
http://www.gesis.org/unser-angebot/studien-planen/zis-ehes/download-ehes/

Internationale & nationale Normen u. Klassifikationen zur Studien- und Datendokumentation

- **Sprache und Geographie**

ISO 639 - Codierung von Namen für Sprachen:

- Beschreibung aller Teilnormen:
http://de.wikipedia.org/wiki/ISO_639
- ISO 639-3 (Einzelsprachen):
http://www.sil.org/iso639-3/codes.asp

ISO 3166 - Codierung bestehender Staaten (ISO 3166-1), staatlicher Untereinheiten (ISO 3166-2) und ehemalige Staaten (ISO 3166-3).

- Listen zu ISO 3166-1; IISO 3166-2 (interner, nicht-kommerzieller Gebrauch; Englisch, Französisch):
  http://www.iso.org/iso/country_codes/iso_3166_code_lists.htm
- Aktualisierung und Newsletter:
  http://www.iso.org/iso/country_codes/updates_on_iso_3166.htm

NUTS - Systematik der europäischen Gebietseinheiten für die Statistik von EUROSTAT:
http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction

Regionalen Standards (2005): Gemeinsame Empfehlung von ADM, ASI und Statistischem Bundesamt:
http://www.gesis.org/fileadmin/upload/dienstleistung/tools_standards/RS2005web.pdf

ISO 19115:2003 - Standard zur Erfassung und Bereitstellung geographischer Information und Dienste:
http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020

- GDI-DE, Geodateninfrastruktur Deutschland, 2008. Deutsche Übersetzung der Metadaten in Annex B ISO 19115:
  http://www.gdi-de.org/download/AK/ISO19115_GermanTranslation_GDIDE.pdf

Bildung und Beruf

ISCED - International Standard Classification of Education der UNESCO (1997; Rev.2011):
http://www.uis.unesco.org/Education/Pages/international-standard-classification-of-education.aspx

- ISCED Mappings 2011:
  http://www.uis.unesco.org/Education/ISCEDMappings/Pages/default.aspx
- ISCO - International Standard Classification of Occupations (ISCO-58, ISCO-68, ISCO-88; ISCO-08) der International Labor Organisation (ILO):
  http://www.ilo.org/public/english/bureau/stat/isco/index.htm

Vergleiche dazu folgende Status- und Prestigemaße und deren Bildung:

- SIOPS - Standard International Occupational Prestige Scale (Treiman, Donald 1971. Occupational Prestige in Comparative Perspective. New York: Academic Press)
- ISEI - International Socio-Economic Index of Occupational Status
  (Ganzeboom, Harry B.G.; De Graaf, Paul; Treiman, Donald J.; (with De Leeuw, Jan). 1992. A Standard International Socio-Economic Index of Occupational Status. Social Science Research (21, 1): 1-56).
- EGP - EGP class scheme (Erikson, Robert; Goldthorpe, John H.; Portocarero, Lucienne, 1979. Intergenerational Class Mobility in Three Western European Societies: England, France and Sweden. British Journal of Sociology (30): 415-451)

  Beschreibung der drei Skalen zum Berufsstatus:
  Ganzeboom, Harry B.G.; Treiman, Donald J., 2003. Three Internationally Standardised Measures for Comparative Research on Occupational Status. In Hoffmeyer-Zlotnik, Jürgen H.P.; Wolf, Christof (Hrsg.). Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables. New York: Kluwer Academic Press: 159-193:

http://home.fsw.vu.nl/hbg.ganzeboom/ismf/..%5CPdf%5C2003-Ganzeboom-Treiman-ComparativeOccupationMeasurement-(chapter-Hoffmeyer).pdf

Quelle der Syntax zur Berechnung der Maße:
http://home.fsw.vu.nl/hbg.ganzeboom/pisa/index.htm

- ESeC - European Socio-economic Classification. User Guide:
  http://www.iser.essex.ac.uk/archives/esec/user-guide

KldB - Klassifikation der Berufe in Deutschland (2010) der Bundesagentur für Arbeit (BA):
http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Berufe/KldB2010/KldB2010-Nav.html

- Systematisches und alphabetisches Verzeichnis:
  http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Berufe/KldB2010/Systematik-Verzeichnisse-Nav.html
- Veröffentlichungen und Dokumentationen:
  http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Berufe/KldB2010/Dokumentationen-Nav.html
- Tabellarische Umsteigeschlüssel:
  http://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Berufe/KldB2010/Umsteigeschluessel-Nav.html
- Paulus, Schweitzer, Wiemer, 2010. Klassifikation der Berufe 2010. Entwicklung und Ergebnis, Methodenbericht der Statistik der Bundesagentur für Arbeit, Nürnberg: 2010:
  http://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Methodenberichte/Arbeitsmarkt-Arbeitsmarktpolitik/Generische-Publikationen/Methodenbericht-Klassifikation-Berufe-2010.pdf

A.4 Technische Metadatenstandards und Initiativen

Technische Metadaten in der Studien- und Datendokumentation der DDI Spezifikationen

DDI – Data Documentation Initiative:

- DDI Alliance Structure:
  http://www.ddialliance.org/alliance/structure
- DDI-Codebook Standard:
  http://www.ddialliance.org/Specification/DDI-Codebook/
- DDI Lifecycle Standard:
  http://www.ddialliance.org/Specification/DDI-Lifecycle/3.1/
- DDI Controlled Vocabularies
  http://www.ddialliance.org/controlled-vocabularies
- DDI Tools
  http://www.ddialliance.org/resources/tools
- DDI at Work
  http://www.ddialliance.org/ddi-at-work

Dublin Core

DCMI - Dublin Core Metadata Initiative:

- Mission and Principles:
  http://dublincore.org/about-us/
- Dublin Core Metadata Element Set, Version 1.1:
  http://dublincore.org/documents/dces/
- ISO 15836:2009:
  http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=52142

Datenzitation, DOI, Metadaten, kontrollierte Vokabulare – da|ra, DataCite, IDF

da|ra, 2012: Registrierungsagentur für sozialwissenschaftliche und wirtschaftswissenschaftliche Forschungsdaten www.da-ra.de mit Informationen und Werkzeugen u. a. zur

- da|ra Policy:
  http://www.da-ra.de/de/ueber-uns/da-ra-policy/policy/
- Service Level Agreement:
  http://www.da-ra.de/de/ueber-uns/da-ra-policy/service-level-agreement/

DataCite, 2012: http://datacite.org/

DataCite, 2011. Metadata Schema for the Publication and Citation of Research Data:
http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

Hausstein, Brigitte 2012: Die Vergabe von DOI-Namen für Sozial- und Wirtschaftsdaten:
http://www.ratswd.de/download/RatSWD_WP_2012/RatSWD_WP_193.pdf

Hausstein et al. 2012. Arbeitsbericht. da|ra-Metadatenschema. Version 2.2.1. September 2012
http://www.da-ra.de/fileadmin/media/da-ra.de/PDFs/Langbeschreibung.pdf

Kontrollierte Vokabulare im da|ra Metadatenschema: ID Person:

- Open Researcher und Contributor ID (ORCID):
  http://www.orcid.org
- International Standard Name Identifier (ISNI):
  http://www.isni.org

Kontrollierte Vokabulare im da|ra Metadatenschema: ID Institution:

- Gemeinsame Körperschaftsdatei (GKD):
  http://www.dnb.de/DE/Standardisierung/Normdaten/GKD/gkd_node.html

Klassifikationen im da|ra Metadatenschema: (Auszug):

- JEL, Journal of Economic Literature Classification System:
  http://www.aeaweb.org/journal/jel_class_system.php
- PSYNDEX Terms:
  http://www.zpid.de/index.php?wahl=products&tuwahl=printed&tuuwahl=psyndexterms
- STW Standard-Thesaurus Wirtschaft:
  http://zbw.eu/wikis/wikisaurus/index.php?n=Main.STW-SystematikUnd-Klassifikation

- Thesaurus Sozialwissenschaften: http://www.gesis.org/unser-angebot/recherchieren/thesauri-und-klassifikationen/thesaurus-sozialwissenschaften/
- Kategorien des GESIS Datenbestandskatalogs http://info1.gesis.org/dbksearch19/Kategorien.htm

IDF - International DOI Foundation: http://www.doi.org/

- DOI® Handbook. Version 5, 2012: http://dx.doi.org/10.1000/186